FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

ADVISORY COMMITTEE FOR CARDIOVASCULAR AND RENAL DRUGS

Discussion of New Drug Application (NDA) 22-449, Binodenoson Injectable, Lypholized Solid 250 Mcg Vial, King Pharmaceuticals Research and Development, Inc., for the Proposed Indication: Short Acting Coronary Vasodilator for Use as an Adjunct to Non-Invasive Myocardial Perfusion Imaging (MPI) Tests to Detect Perfusion Abnormalities in Patients with Known or Suspected Coronary Artery Disease (CAD)

TUESDAY, JULY 28, 2009

8:00 a.m. to 4:45 p.m.

Hilton Washington, D.C./Silver Spring

8727 Colesville Road

Silver Spring, Maryland

CARDIOVASCULAR AND RENAL DRUGS ADVISORY COMMITTEE MEMBERS (Voting)

Robert A. Harrington, M.D., F.A.C.C. (Chair)
Professor of Medicine, Division of Cardiology
Department of Medicine
Duke University Medical School
Duke Clinical Research Institute
2400 Pratt Street, NP 7007
Room 0311, Terrace Level
Durham, North Carolina 27705

Henry R. Black, M.D.
Clinical Professor of Internal Medicine
Department of Internal Medicine
New York University School of Medicine
550 First Avenue
New York, New York 10016

Jonathan L. Halperin, M.D.
Professor of Medicine, Cardiology
Mount Sinai Medical Center
Fifth Avenue at 100th Street
New York, New York 10029

Sanjay Kaul, M.D.
Director, Cardiovascular Diseases
Fellowship Training Program
Cedars-Sinai Heart Institute
Division of Cardiology, Room 5536 S. Tower
8700 Beverly Blvd
Los Angeles, California 90048

Mori J. Krantz, M.D., F.A.C.C.
Associate Professor
University of Colorado/Cardiology, Denver Health
Director, CV Prevention, & ECG Core Lab
Colorado Prevention Center
789 Sherman St. Suite 200
Denver, Colorado 80203

Darren K. McGuire, M.D., M.H.Sc.
Associate Professor of Medicine
University of Texas Southwestern
Medical Center
5323 Harry Hines Boulevard
St. Paul Hospital, Suite HA9.133
Dallas, Texas 75390

James D. Neaton, Ph.D.
Professor of Biostatistics
Division of Biostatistics
Coordinating Centers for Biometric Research
University of Minnesota
School of Public Health
2221 University Avenue S.E., Suite 200
Minneapolis, Minnesota 55414

Emil P. Paganini, M.D., F.A.C.P., F.R.C.P.
Critical Care Nephrology Consulting
10427 Mayfield Road
Chesterland, Ohio 44026

(Non-Voting)

Jonathan C. Fox, M.D., Ph.D., F.A.C.C.

(Industry Representative)

Vice President, Clinical Therapeutic Area

Cardiovascular and Gastrointestinal Diseases

AstraZeneca LP

PO Box 15437

Wilmington, Delaware 19850

(Temporary Voting)

Frank M. Bengel, M.D.

Associate Professor of Radiology and Medicine

Director of Cardiovascular Nuclear Medicine

Johns Hopkins Medical Institutions

Baltimore, MD 21205

Lyle D. Bromeling, Ph.D.

Director, Broemeling and Associates Inc.

Adjunct Professor, Whitworth University

Medical Lake, Washington 99022

Peter Conti, M.D., Ph.D.
Professor of Radiology
USC Keck School of Medicine
1510 San Pablo Street, Suite 350
Los Angeles, California 90033

Michael Domanski, M.D., Ph.D.
Chief Atherothrombosis and Coronary
Artery Disease Branch
NHLBI/NIH

John M. Flack, M.D., M.P.H., F.A.H.A.
Professor of Medicine and Physiology
Chair, Department of Medicine
Chief, Division of Translational Research
and Clinical Epidemiology
Wayne State University
4201 St. Antoine, Suite 2E
Detroit, Michigan 48201

Sebastian G. Schneeweiss, M.D., Sc.D.
Associate Professor of Medicine and
Epidemiology, Harvard Medical School
Vice Chief, Division of Pharmacoepidemiology,
Dept. of Medicine, Brigham & Women's Hospital
1620 Tremont Street, Suite 3030
Boston, Massachusetts 02120

Temporary Voting)

James L. Tatum, M.D.
Associate Director DCTD
Chief - Molecular Imaging Branch
CIP/DCTD/NCI/NIH

Neil J. Weissman, M.D.
Director, Cardiac Ultrasound &
Ultrasound Core Labs
President, MedStar Research Institute
Professor of Medicine, Georgetown University
100 Irving Street, N.W.
Washington, District of Columbia 20010

I N D E X

P R O C E E D I N G S

- - - - -

DR. HARRINGTON: Why don't we go ahead and get started. It's right at 8:00. My name is Bob Harrington. I'm a cardiologist at Duke University, and I'll chair the meeting today.

I'm going to read an opening statement that we're required to read, and then I'd like to go around and have the advisory panel introduce themselves before I turn it over to Elaine to read the conflict of interest statement.

For topics such as those being discussed at today's meeting, there are often a variety of opinions, some of which are quite strongly held. Our goal is that today's meeting will be a fair and open forum for the discussion of these issues, and that individuals can express their views without interruption. Thus, as a gentle reminder, individuals will be allowed to speak into the record only if recognized by the chair. We look forward to a productive meeting.

In the spirit of the Federal Advisory

Committee Act and the Government in the Sunshine Act, we ask that the advisory committee members take care that their conversations about the topic at hand take place in the open forum of the meeting. We are aware that the members of the media are anxious to speak with the FDA about these proceedings. However, FDA will refrain from discussing the details of this meeting with the media until its conclusion. Also, the committee is reminded to please refrain from discussing the meeting topic during breaks or lunch.

So, Dr. Fox, why don't we start with you, and if we could go around the table, introduce yourself and your area of expertise and your institution.

DR. FOX: My name is Jonathan Fox. I'm a cardiologist in clinical development with AstraZeneca, and I'm the industry representative to the committee.

DR. CONTI: I'm Peter Conti. I'm a professor of radiology and nuclear medicine at USC in Los Angeles.

DR. WEISSMAN: Neil Weissman. I'm a cardiologist at Washington Hospital Center, MedStar, and professor of medicine at Georgetown.

DR. FLACK: John Flack. I'm a professor of medicine and physiology, cardiovascular epidemiologist, hypertension specialist, at Wayne State University in Detroit.

DR. SCHNEEWEISS: Sebastian Schneeweiss. I'm a general internist and pharmacoepidemiologist. I'm an associate professor of medicine in epidemiology at Harvard Medical School.

DR. TATUM: I'm Jim Tatum. My background in radiology, nuclear medicine, and nuclear cardiology. I'm currently associate director of the National Cancer Institute.

DR. BROMELING: I'm Lyle Bromeling, retired professor of biostatistics from M.D. Anderson Cancer Center.

DR. KAUL: Sanjay Kaul. I'm a cardiologist at Cedars Sinai Medical Center in Los Angeles.

DR. KRANTZ: Good morning. Mori Krantz, cardiologist, University of Colorado in Denver.

DR. PAGANINI: Emil Paganini, private nephrologist, former section head of Critical Care

Nephrology, Cleveland Clinic Foundation, Cleveland, Ohio.

MS. FERGUSON: Elaine Ferguson, designated federal official.

DR. BLACK: I'm Henry Black. I'm a clinical professor of internal medicine at New York University, a hypertension specialist.

DR. HALPERIN: Good morning. I'm Jonathan Halperin, a cardiologist at the Mount Sinai Medical Center in New York, where I am professor of medicine in cardiology.

DR. McGUIRE: Darren McGuire, University of Texas Southwestern Medical Center at Dallas, general cardiology.

DR. NEATON: Jim Neaton. I'm professor of biostatistics, University of Minnesota.

DR. BENGEL: Frank Bengel, radiologist and nuclear cardiologist, Johns Hopkins University in Baltimore.

DR. MARZELLA: I'm Lou Marzella in the Division of Medical Imaging at FDA.

MR. LEVENSON: I'm Mark Levenson, a

statistical reviewer at FDA.

DR. REEVES: Hi. I'm Duane Reeves, director of the Division of Imaging and Hematology at the FDA.

DR. UNGER: Good morning. I'm Ellis Unger, a cardiologist, deputy director of Office of Drug Evaluation I, FDA.

MS. FERGUSON: The Food and Drug Administration, FDA, is convening today's meeting of the Cardiovascular and Renal Drugs Advisory Committee under the authority of the Federal Advisory Committee Act, FACA, of 1972.

With the exception of the industry representative, all members and temporary voting members of the committee are special government employees, SGEs, or regular federal employees from other agencies, and are subject to federal conflict of interest laws and regulations.

The following information on the status of this committee's compliance with federal ethics and conflict of interest laws covered by, but not limited to, those found at 18 USC Section 208 and Section 712 of the Federal Food, Drug, and Cosmetics Act, FD&C

Act, is being provided to participants in today's meeting and to the public.

FDA has determined that members and temporary voting members of this committee are in compliance with the federal ethics and conflict of interest laws under 18 USC Section 208. Congress has authorized FDA to grant waivers to special government employees and regular federal employees who have potential financial conflicts when it is determined that the agency's need for a particular individual's services outweighs his or her potential financial conflict of interest.

Under Section 712 of the FD&C Act, Congress has authorized FDA to grant waivers to special government employees and regular federal employees with potential financial conflicts when necessary to afford the committee essential expertise.

Related to the discussions of today's meeting, members and temporary voting members of this committee have been screened for potential financial conflicts of interest of their own, as well as those imputed to them, including those of their spouses or minor children, and, for purposes of 18 USC Section 208, their employers.

These interests may include investments, consulting, expert witness testimony, contracts, grants, CRADAs, teaching, speaking, writing, patents and royalties, and primary employment.

Today's agenda involves discussion of King Pharmaceuticals' New Drug Application for binodenoson injectable, lypholized solid, 250 microgram vial, for the proposed indication: short acting coronary vasodilator for use as an adjunct to noninvasive myocardial perfusion imaging tests to detect perfusion abnormalities in patients with known or suspected coronary artery disease.

This topic is a particular matter involving specific parties. Based on the agenda for today's meeting and all financial interests reported by the committee members and temporary voting members, no conflict of interest waivers have been issued in connection with this meeting.

To ensure transparency, we encourage all standing committee members and temporary voting members to disclose any public statements that they have made concerning the product at issue.

With respect to the FDA's invited industry representative, we would like to disclose that Dr. Jonathan Fox is participating in this meeting as a nonvoting industry representative, acting on behalf of the regulated industry.

Dr. Fox's role at this meeting is to represent industry in general and not any particular company. Dr. Fox is employed by AstraZeneca.

We would like to remind members and temporary voting members that if the discussions involve any other products or firms not already on the agenda for which an FDA participant has a personal or imputed financial interest, the participants need to exclude themselves from such involvement, and their exclusion will be noted for the record.

FDA encourages all the other participants to advise the committee of any financial relationships that they may have with any firms at issue.

And now I would like to identify the FDA press contact, Karen Riley, and also Brigit Henig.

Thank you very much.

DR. HARRINGTON: Thanks, Elaine.

Before I turn it over to Dr. Rieves, Dr. Domanski, if you could just introduce yourself for the record.

DR. DOMANSKI: Mike Domanski. I'm an interventional cardiologist, National Heart, Lung, and Blood Institute.

DR. HARRINGTON: Terrific. Thanks, Mike.

We're going to open with a statement from the FDA by Dr. Rieves, the director of the division.

DR. RIEVES: Good morning. I have a few prepared remarks to set the stage for today's discussion.

On behalf of our Imaging and Hematology Review Division, we welcome you to our discussion of a New Drug Application for CorVue, which is the proposed trade name for binodenoson injection.

CorVue is a pharmacologic stress agent, that is, a drug which is somewhat intended to mimic the effect of exercise stress upon the heart and coronary circulation.

As listed here, and as Elaine noted, the drug is specifically proposed to be indicated as a short acting coronary vasodilator for use as an adjunct to noninvasive myocardial perfusion imaging, or MPI, tests to detect perfusion abnormalities in patients with known or suspected coronary artery disease.

This proposal makes it clear that the drug is to be used as an adjunct in diagnostic imaging. Hence, the main Phase 3 study outcomes were pictures of cardiac radionuclide uptake before and after administration of the drug.

FDA regulations and guidance documents are relatively specific in the efficacy expectations for diagnostic imaging agents. The establishment of performance characteristics is generally regarded as the optimal goal for a new imaging agent, that is, establishment of the agent's diagnostic sensitivity and specificity based upon comparison of the images to a standard of truth; for example, comparison of radionuclide-based images to coronary arteriographic images, an accepted standard of truth.

Alternatively, a new agent's efficacy may be established by confirmation of agreement between an accepted reference test's images and the new agent's images. The consequence of this type of comparison is that the two agents, the new and the reference agent, would be regarded as diagnostically interchangeable.

Regarding agreement between a new and reference test, our guidance documents note that as an alternative to the establishment of performance characteristics, similarity between a new test agent and a reference product can also be shown by demonstrating that both agents consistently give identical results.

Subsequent text elaborates a bit more by stating that high agreement between a new test product and a reference product can support a claim that the new test is an acceptable alternative to the reference product. So what is high agreement?

In essence, we regard high agreement as demonstration that the reference product images are the same as those of the new agent's images, or, at a minimum, the images are the same with respect to clinically important image aspects, such as the extent of cardiac perfusion defects, and that the images are of high technical quality.

Why is the concept of high agreement so important? As previously mentioned, diagnostic imaging agents are best characterized by their performance characteristics when compared to a truth standard.

When a new agent's images are solely compared to reference test images, the new agent's performance characteristics are inferred to be the same as those of the reference test. And commonly, clinical studies based upon agreement do not contain features that allow direct verification of the reference agent's performance within the clinical studies.

Hence, agreement between a new agent and a reference agent could be clinically meaningless if the images were of poor quality or if clinically meaningless aspects of the images were compared.

To date, the FDA has approved three drugs specifically for use in pharmacologic stress:

Dipyridamole, approved in 1990, had performance characteristics established using coronary arteriography as a truth standard;

Adenosine, approved in 1995, had efficacy established using a coronary arteriographic truth standard as well as comparison to exercise stress images;

Regadenoson, approved last year, was the first agent to have its efficacy based entirely upon comparison to a reference test, adenosine-based images.

All three of these agents are approved for use among patients who are unable to exercise adequately.

Last year's approval was particularly illustrative in that the confirmatory studies consisted of two Phase 3 clinical studies, where the primary endpoints compared concordance between the new and reference agent images. The study results were consistent between the two studies, perhaps somewhat related to the study designs that quantified the test/retest variability of the reference agent.

In these studies, patients all underwent two sets of myocardial perfusion imaging, with the first imaging performed with adenosine. Subsequently, patients were randomized to either the new agent or to repeat imaging with adenosine. Hence, the new agent's images could not only be directly compared to the reference test images, but the reference test variability could also be incorporated into the comparisons. This type of study design was not used in the binodenoson Phase 3 studies.

One of the challenges with the binodenoson efficacy data set pertains to the change in endpoints for two of the Phase 3 studies. Two years ago, prior to unblinding of the image data, the FDA was requested to comment upon the proposed primary endpoint revisions. We did not agree with the proposed revision, and cited here are some quotes that illustrate our general perspective.

We noted that: "These proposed alterations are fundamental alterations of statistical, clinical, and technical assumptions." We went on to say that: "We suggest that you retain the primary endpoint and statistical methodology, as currently described, but modify the protocols and analytical plans to include pre-specified exploratory analyses of the primary endpoint."

We further noted, as we always try to note to sponsors, that: "In this regard, we anticipate the review of the totality of findings, primary, secondary, and exploratory endpoint results, in assessing efficacy."

Given these challenges, our review team has brought binodenoson to this committee for largely a single purpose, which is articulated here as the question:

Do the Phase 3 study results establish high binodenoson and adenosine MPI agreement? In particular, the data have been challenging in that the primary endpoint in the first Phase 3 study, a comparison of concordance, was not achieved.

Subsequently, the originally stated primary endpoints for the other two Phase 3 studies were changed to comparisons of an average perfusion defect score, referred to as the summed difference score, or

SDS.

We have no regulatory precedent for the use of SDS scores in this matter, and we are also unclear of the clinical meaningfulness of incremental changes in these scores. Overall, the original primary endpoints were not achieved in the Phase 3 studies, while the revised primary endpoints were achieved. Inconsistency in these results has raised questions as to the extent of agreement between the tested agents.

Lastly, I want to emphasize that we are not coming to this committee with a finalized, complete review of the New Drug Application. Indeed, this discussion today is a component of our review process, where we are looking forward to your perspectives on the data as you understand it, such that you can help us all refine our final review focus.

Thank you for your help. And, Mr. Chairman, I return the podium to your direction.

DR. HARRINGTON: Thank you, Dr. Rieves.

So just as a point of order, we now have approximately an hour and 40 minutes or so, 45 minutes, before a break. As we usually do at these, and I think most of you know this, we'll let the sponsor go through their presentation and then have a period of questions after that. Of course, if there's a burning question that you just need a point of clarification, just indicate that to me so that we'll try to get that squeezed in earlier.

I think we're going to have a lot of time for questions throughout the day. So write your questions down and we'll start that right after the break.

So with that as an introduction, I'll turn it over to Dr. Carter from the sponsor to make introductions.

DR. CARTER: Mr. Chairman, members of the Cardiovascular and Renal Advisory Committee, ladies and gentlemen, good morning. My name is Eric Carter, and I'm the chief science officer at King Pharmaceuticals, and I will coordinate our presentations this morning as well as the Q&A session.

As you've just heard from Dr. Rieves, FDA has convened this meeting to provide advice concerning certain concerns the agency has with the diagnostic efficacy data of our NDA for binodenoson. And as you've also heard from Dr. Rieves, binodenoson is an injectable short acting coronary vasodilator for use as an adjunct to noninvasive myocardial perfusion imaging tests to detect perfusion abnormalities in patients with known or suspected coronary artery disease. And the proposed indication is again shown on this slide.

The FDA briefing document and in Dr. Rieves' introductory comment, you will have noted that FDA has a number of concerns, then, regarding the approach that we've taken in developing binodenoson and with some of our results. One of the concerns is that based on the data that became available during development, we amended the primary endpoint of our pivotal Phase 3 program during the conduct of the trials.

We will show you why this was done and demonstrate that the statistical analysis plan was amended in full compliance with ICH guidelines, and why the results of the two pivotal trials can be regarded as confirmatory.

FDA is also concerned that the images obtained from binodenoson and adenosine do not sufficiently agree, and as a result, these two agents cannot be claimed to be diagnostically interchangeable.

We will show you data using multiple approaches that consistently indicate that binodenoson, by increasing coronary blood flow to the same extent as adenosine, has similar diagnostic performance characteristics, and therefore, that it's equivalent to adenosine.

As a key part of our presentation, we will show you why it was appropriate to amend the primary efficacy analysis, and why amending this to an analysis based on clinical equivalence is valid for this type of comparison between two imaging agents and for the population intended to be exposed to binodenoson following approval.

As requested by FDA, efficacy will be demonstrated based on the positive outcome of several endpoints, both primary and secondary, in other words, based on the totality of the data.

The rationale behind the development of binodenoson as a selective A_{2A} receptor agonist was to improve the tolerability and safety profile of myocardial perfusion imaging products. Although we recognize that addressing the safety profile of binodenoson is not the primary objective of today's meeting, we will show that binodenoson on the whole is associated with fewer and less severe adverse events than adenosine.

We therefore believe that, overall, binodenoson has a more favorable benefit-to-risk profile than adenosine, and that it's important to consider this in the context of a new product review and approval.

Given the high prevalence of coronary artery disease in the U.S., the need for accurate and risk stratification is essential for clinicians to make appropriate treatment decisions. It's therefore not surprising that more than 9 million noninvasive myocardial perfusion imaging procedures are performed in the U.S. each year, and almost one-half of these employ pharmacologic stress to generate important diagnostic information for clinical decision-making.

So binodenoson is a selective A_{2A} receptor agonist specifically developed as a pharmacologic stress agent for myocardial perfusion imaging studies. And as a reminder, pharmacologic stress testing is utilized when coronary blood flow cannot be sufficiently increased by exercise, and under these circumstances, coronary arterial vasodilatation in conjunction with imaging using the uptake of a radioisotope and computed tomography, the so-called SPECT imaging.

This diagnostic method is now commonly utilized in the evaluation of known or suspected ischemic heart disease as a gateway to angiography. Thus, a very important objective of these diagnostic tests is to identify those patients that have a low likelihood of clinically significant reversible ischemia in order to avoid having to expose them to angiography.

In this context, then, binodenoson was developed because of its selectivity for the A_{2A} receptor and the expectation that, as a result, it would provide equivalent coronary hyperemia to adenosine, and therefore enable equivalent clinical decision-making, but would be safer and better tolerated.

As I've mentioned, adenosine is a suitable comparator because it's the most widely used stress agent in the U.S. Dipyridamole increases the concentration of adenosine, and because of this it's also used.

So whether given directly or indirectly, adenosine binds to the A_{2A} receptor to cause coronary vasodilatation, decrease resistance, and thereby increase coronary flow.

Unfortunately, adenosine also exerts its pharmacological effect through the activation of the other adenosine receptors, as shown on the slide. This is unfortunate for a pharmacologic stress agent because these other receptors mediate undesirable side effects that include AV block, chest pain, flushing, dyspnea, and bronchospasm, side effects which when they occur during a procedure are of obvious concern to patients and clinicians alike.

In fact, side effects occur frequently with a pharmacological stress agent. These data from an adenosine registry of almost 10,000 patients demonstrates this. As you can see, overall, approximately 90 percent of patients reported side effects, and as highlighted in the upper red box, about a third complained of either flushing, shortness of breath, or chest pain.

In this particular registry, shown in the lower red box, about 7 percent of patients had AV block, with about 5 percent experiencing second- or third-degree block. And these data were used to define the adverse events of special interest that were then prospectively measured as part of the clinical development plan for binodenoson.

A pharmacologic stress agent such as binodenoson, by being more selective for A_{2A}, theoretically enables vasodilatation and hyperemia but with fewer side effects, particularly AV block, flushing, chest pain, and dyspnea, than nonselective agents such as dipyridamole or adenosine. And indeed, as you'll see in our presentation, binodenoson was found to be selective for the A_{2A} receptor, and this resulted in coronary hyperemia similar to adenosine, but was better tolerated by patients.

Before moving on to the core of our presentation, I'd like to point out for the committee some key elements that occurred during the Phase 3 development program.

Enrollment for Study 301 started in December 2003, and Study 302 in February of 2004. These studies were intended to be the two primary efficacy studies. The design of the trials occurred prior to the release of the final guidance, which occurred in June of 2004, but did generally conform to the FDA draft documents that were available at the time. The final June 2004 guidance formalized design principles recommended by FDA for the development of imaging agents such as binodenoson.

The results of Study 301 became available in February of 2005. The study failed to meet its primary endpoint. Interrogation of the results, together with a review of information from the regadenoson clinical development program that had become publicly available, led us to recognize that better methods were needed to address the sources of variability that are associated with myocardial perfusion imaging studies.

Simply put, we used a statistical approach that was based on the limited data available at the time, but which proved to be inappropriate when challenged by a much larger body of data with images not collected simultaneously and according to the FDA guidance document.

As a result, our pre-specified efficacy analysis for estimating concordance using the kappa statistic at a high threshold, was found to be inadequate to assess the agreement between two sets of pharmacologic stress imaging procedures performed on the same patient. And so, the efficiency analysis was changed to a clinical equivalence analysis. This amended statistical methodology was then prospectively applied to two confirmatory primary efficacy studies, Study 302 and Study 305, and prior to unblinding.

We'll describe our rationale in much greater detail today, and will demonstrate why this approach is valid, rigorous, and therefore why we believe that it shouldn't be considered as exploratory or supportive.

For Study 302, the statistical analysis plan and protocol was amended prior to database lock and unblinding. Meanwhile, in October of 2005, we had started enrolling patients in Study 305, the confirmatory study of Study 302. But here we had design elements that were not consistent with the final guidance document. Importantly, an adenosine/adenosine treatment arm was included to allow estimation of test/retest of method-to-method variability with the same agent in the same patient.

As you can see, the statistical analysis plan and protocol for 305 was amended prior to the images being read and unblinded. All other aspects of the trials -- the patient population, inclusion and exclusion criteria, the process for collecting, reading, and interpreting data, and so on -- was unchanged.

Amending the statistical analysis plan and the protocols was done in full accordance with ICH guidelines. In fact, a subsequent independent audit proved that the blind had been maintained on the images read for the efficacy analysis.

Our presentation therefore will provide details on why amending the primary endpoint to a clinical equivalence analysis was sensible, rational, and valid. We'll also focus on the clinical relevance of angiography data collected on about 15 percent of the Phase 3 trial population.

And we will show that the totality of the data demonstrates that binodenoson, the test agent, and adenosine, the reference, are diagnostically interchangeable. You will see data demonstrating that selectivity for the A_{2A} receptor confers improved tolerability for binodenoson. And finally, we'll show you that the benefit-to-risk profile for binodenoson is favorable relative to adenosine.

Dr. James Udelson will now present the clinical development program in detail. Dr. Udelson is chief of the division of cardiology at Tufts Medical Center. He's an expert in the field of cardiovascular imaging and has been involved with the binodenoson development program from very early on.

Dr. Udelson has guided and advised us. He's very familiar with all aspects of the program. And it's fitting that he should present the efficacy and safety data.

Since a discussion on the statistical treatment of the efficacy data is central to why we're here today, we've asked Dr. Lisa LaVange to hone down on the key statistical considerations that underpin the efficacy data.

Dr. LaVange is professor of biostatistics at the University of North Carolina in Chapel Hill, and director of the Collaborative Studies Coordinating Center. Dr. LaVange has also been a consultant on this project for a considerable period of time, and has provided us much appreciated input and direction.

Dr. Udelson will then present the Phase 3 efficacy and safety results, and I will end with some concluding remarks.

Dr. Udelson.

DR. UDELSON: Thank you very much. Before I start, I'd like to just state clearly that some of you know I'm a special government employee and was a voting member of this panel back in February during the prasugrel meeting. But I've received permission from FDA to appear here today as a presenter based on established FDA criteria in communications with Ms. Ferguson.

So what I'd like to do in the next few minutes is give an overview of the clinical development program, and then talk in some detail about the Phase 2 studies on coronary hyperemia, take a little bit of a detour to discuss SPECT imaging analytic methodology, which is so central to the understanding of the whole program, and then talk about the dose identification study, and then begin a discussion of the Phase 3 pivotal program.

So this slide is an overview of the entire clinical development program for binodenoson: the early studies on PK and PD; initial safety and dose-finding studies to narrow down a wide dose range into a smaller range; selecting an IV dosing regimen for optimal coronary hyperemia; assessing the potential for bronchoconstriction, or lack thereof, actually;

evaluation of some imaging parameters; the reversibility with aminophylline, which is clinically important; ultimately leading in Phase 2 to what we call Study 206, which was the study designed to find a dose to move on to the Phase 3 program, and then ultimately, as Dr. Carter mentioned, three active control, double-blind, double-dummy, multicenter trials, 301, 302, and 305.

So let me discuss the coronary hyperemia studies, which is really, of course, the basis for how an adenosine A_{2A} receptor agonist works. This is a Doppler flow wire recording from a patient in the 202 study, which was published a few years ago -- a couple of years ago in the American Journal of Cardiology.

In this study, patients were in a cath lab. At least one of their coronary arteries was normal or near normal. And within that artery, they had an injection of intra-coronary adenosine to create a reference for an increase in coronary blood flow velocity, which is what is shown here on the Y axis over time.

So here are three doses of intra-coronary adenosine. As you can see, a rapid and very transient increase in blood flow velocity. And at this point, binodenoson, in this particular case at a dose of 1.5 mics per kilogram as a 30-second bolus, was given, and this ultimately, as you know, went on to the Phase 3 program as the dose that was used. And you see a rapid increase in coronary blood flow velocity to similar levels as adenosine, lasting clearly long enough for extraction of a radioisotope, which is what is needed, and a longer half life than intra-coronary adenosine.

This slide summarizes the Study 202 data. And there were multiple doses of binodenoson used. Again, we'll focus a little bit on the 1.5 mic per kilo dose. Intra-coronary adenosine was the reference standard here.

This is the percent of coronary blood flow velocity reserve achieved, in other words, the percent of the coronary blood flow velocity reserve of adenosine. And if we just focus in this box here on the 1.5 dose, you can see that near 100 percent of the coronary blood flow velocity reserve was achieved, compared to adenosine, with this dose of binodenoson. And at the bottom of the slide, the peak coronary blood flow velocity that was observed was similar -- this is the mean and the standard deviation -- similar to that observed with adenosine.

Note the wide range here. And note the wide range also, by the way, with intra-coronary adenosine. So there is some variability in the coronary hyperemic responses with binodenoson, but also with the drug that's considered, in this particular study and others, the gold standard for increase in coronary blood flow.

So at that point the dose range had been narrowed. It seemed that coronary hyperemia occurred to a similar degree as adenosine. So as I mentioned, I'd like to take a few minutes to talk about the assessment of SPECT myocardial perfusion images, again because this is very central to much of the discussion, and understanding some different scores, et cetera.

Per FDA guidance and per professional society guidelines, there's a visual evaluation of myocardial perfusion images in 17 standardized myocardial segments in a model of the myocardium for both the rest and the stress images. And you score these segments on a scale of 0 to 4, where 0 is normal uptake of the tracer and 4 is a severe defect; 1, 2, and 3 are gradations in between.

Here is the standardized 17-segment model that is supported by the American College of Cardiology, American Heart Association, and the American Society of Nuclear Cardiology, published a few years ago in a paper in circulation. And one of the panelists today, Dr. Weissman, was the second author on this paper.

So the myocardium is segmented into 17 segments representing the different vascular territories. And then each of these segments is scored on a scale of 0 to 4 for the rest images, and again for the stress images.

Then you add up the scores. You add up all of the 17 segmental scores at rest, and you come up with what's called the summed rest score, SRS. And this represents the extent -- in other words, how many segments are abnormal -- and the severity -- how abnormal each segment is when you add it all up -- of the resting perfusion abnormality. And the clinical relevance here; in general, this represents the extent of infarction.

If you add up the scores from the stress image, you get the summed stress score, the SSS. This represents the extent and severity of the stress perfusion abnormality, the clinical relevance being both the extent of infarction and the extent of inducible ischemia. And then when you subtract the summed rest from the summed stress score, you get what's known as the summed difference score, or SDS. And this is what we'll be talking about a lot today, the SSS minus the SRS.

The clinical relevance is it represents the extent and severity of inducible ischemia -- again, extent because it's the number of segments that are abnormal that are added up, and the severity because each segment can be scored from 0 to 4. So it's one number that represents sort of a global extent and severity of inducible ischemia as you might see during exercise or during a pharmacologic stress agent.

So here's an example of what a reader might see in a core lab. So these first three columns are short axis images at the basal, mid, and apical portion of the myocardium. Stress images are on top. The corresponding rest tomogram is on the bottom.

And this is a vertical long axis image, which should look like a sideways U here -- anterior wall, apex, inferior wall, and in the short axis images, should look like a yellowish doughnut -- anterior wall, lateral wall, inferior wall, and septum.

So a reader in a core lab might then look at this and visually score the segments thus. So these look pretty normal, so you would score a 0. This is fairly severely but not terribly severely abnormal, so I'll give that a 3. This looks like a 2. This is very severe; I'll call that a 4. And then each segment of the 17 is scored at stress and at rest.

You can also create segmental difference scores. So if you subtract 4 from 4, you get a summed difference score of 0 for the apex. So there's no ischemia; it's just an infarction. The lateral wall here goes from darker yellow to brighter yellow, from a 2 to a 0, so that's an area of inducible ischemia, as well as over here in the inferolateral wall.

So in seeing this, you can begin to understand that there's some variability associated with this, even if you have expert readers who do this a lot. I mean, it's a human eyeball endeavor, as it were.

So in this particular example, the summed stress score is 24 when you add up these, the summed rest score is 15, and the difference between the two, which would represent the extent of ischemia in this particular scan, is 9, sort of the global extent of ischemia.

Now, nuclear cardiology is somewhat unique among all of the imaging modalities in that there is widely applied quantitative analysis programs that are used in almost every laboratory in the country that are validated and FDA-approved.

For the purposes of this study, we used a program called the 4D-MSPECT study that was developed by Dr. Ficaro, who is here today. And here are the images. And essentially, the images are collapsed into a two-dimensional plot, and then these 17 segments are overlaid on top of the two-dimensional summary of the three-dimensional data. And using an internal standard, the computer scores the segments here at the bottom, as you can see, 4, 3, 2, 1, 0, and then sums them up.

So in this example, which is different than the previous slide, the summed stress score is 28, the summed rest score is 5, and the difference is 23, representing substantial extent and severity of ischemia. So later on today I'll show you some data using just the computer-based analysis as well as the human visual analysis.

Now, these scores, the summed stress score, the summed difference score, have been clinically validated, in a sense, because there's an enormous literature within the cardiology and nuclear cardiology literature looking at their prognostic value related to outcomes.

And this is an example of one such study, looking at the summed stress score. There's over a thousand patients who were referred for an adenosine SPECT study, two years of follow-up for heart events, cardiac death, or myocardial infarction, 11 percent heart event rate over the two years.

There were two groups of patients. On the left, low likelihood, pretest likelihood, of coronary disease; on your right, intermediate to high pretest likelihood. The scans, normal, summed stress score 0 to 3 in green, gold is mildly abnormal, and purple is moderate to severely abnormal, as defined by these numbers that you see.

And within both pretest probability categories, there is risk stratification information. In other words, there's a difference in the predicted outcome rate, the rate of cardiac death or myocardial infarction, across the scanned categories within each pretest likelihood group. And the asterisks here means the P value is less than 0.001 for differences across the scanned categories within each likelihood group.

This was published many years ago, but there are many, many studies, literally hundreds of studies in the literature, that look like this, validating the use -- clinical validation for risk stratification of the summed stress score for predicting outcome event risk.

Now, there's also a literature on using the information, that clinicians use the information to refer patients to catheterization. So these are data from the same study, looking now on the Y axis at the rate of referral to catheterization based on the imaging results. And so these are the MPI results. This is the rate of referral. So as the scan gets more abnormal, clinicians refer the patients to catheterization at a higher rate. Now, that's fairly intuitive, I would say, but it's established in literature, at least, that clinicians respond to the results in this manner. And in fact, in this particular study, in a multiple logistic regression model, only the summed difference score, the extent of ischemia, was an independent predictor of referral to catheterization. And anybody in cardiology would not be shocked by that. The more ischemia, the more likely you are to refer the patient to catheterization.

So now back to the development program. So the Study 206 was done to select a dose that would move on to the Phase 3 trial. So the objective was to select the optimal binodenoson dosing regimen to move on to the Phase 3 program, with the idea being that the optimal dose would be a balance, would provide a balance, of the most concordant SPECT images with adenosine regarding the extent and severity of reversible defects, with the most favorable safety profile, which really means a reduction in side effects because, remember, the reason to develop this category of drugs at all, as Dr. Carter mentioned, is the selectivity at the A_{2A} receptor.

So if you have sufficient coronary hyperemia, which it seemed to by the 202 results, the idea is you should get similar images but with fewer side effects compared to adenosine. And we wanted to find the dose that optimally balanced that.

So the 206 study was designed as such: eligible patients who were targeted to have 10 percent high pretest likelihood of coronary disease and 90 percent with known coronary disease. The reason here is that we wanted to see a good amount of ischemia so we would have a lot of SDS to work with, as it were, within four dose groups.

All of the patients had both an adenosine SPECT study and a binodenoson study. They were randomized to a sequence, either adenosine first, bino second, or bino first, adenosine second, in a double-blinded manner, over two to seven days between the procedures.

Now, in this trial and throughout the development program in Phase 3, extensive efforts were made to minimize variability with extensive training of sites, investigators, nuclear technologists, et cetera. And in fact, the sites were instructed to standardize the acquisition as much as possible between the first and the second imaging session, to use the same camera, the same imaging protocol, the same isotopes, the same doses, the same acquisition times, imaging times after dosing, about the same time of day. And it was recommended in these stable patients that background medications were held on the day of the testing until after the testing was over unless the PI did not feel comfortable with that. And all of this was tracked very carefully and monitored.

Now, for the reading in this particular study, the reading was done in a blinded core lab. The readers were shown both images from a patient side by side. So if it's patient No. 22, let's say, this might have been their bino image. This might have been their adeno image. The reader didn't know. These are the electronic case report forms with the 17-segment model.

Now, it's important to note the side-by-side reading is not in keeping with FDA guidance for image analysis in pivotal clinical trials. This was a dose-finding trial, where we were trying to find the best dose or, in essence, the most superior dose to move on to Phase 3.

So the idea here, the rationale for the side-by-side read, was that it would minimize the read-to-read variability so that we could see a signal of concordance without the noise in a modest-sized trial. So we thought a lot about this. And again, the readers were blinded to the agent, but the study ultimately was really designed for dose-finding. And we'll get back to this point a little bit later.

So here are the results, in essence, the efficacy results. There were four dosing groups in 206: .5 bolus; 1 mic per kilo bolus; 1.5 bolus; and .5 mic per kilo, 3-minute infusion, because this seemed to be efficacious in prior studies as well.

We used three different metrics of efficacy in this study: the percent categorical agreement within SDS categories, which we'll talk more about in a little bit; a weighted kappa statistic across those categories; and using SDS as a more continuous variable, a coefficient of determination.

I think, as you can see here in the red box, the steering committee, in looking at all of these data as well as the side effect data, thought that the 1.5 mic per kilo intravenous bolus dose seemed to have the best concordance with adenosine.

The categorical agreement was high, a bit higher than the other groups; the weighted kappa statistic seemed to be higher than the others; and the coefficient of determination was higher as well.

These data were published a few years ago in circulation, and what I don't have to show you at the moment is in this study, the side effects were reduced by about 50 percent in the 1.5 dose. So that seemed to be a good dose to move ahead on to the Phase 3 trial.

This is how the weighted kappa statistic was calculated in this and the subsequent trials we'll talk about. There were four categories of summed difference score, or SDS, from normal, which really means nonischemic score of 0 to 1, mildly ischemia, 2 to 4, moderately ischemic, 5 to 8, and then more severely ischemic, greater than 8. And these are categories that are based on studies in the literature, adenosine on top, binodenoson studies along the side, with the purple boxes being the exact categorical agreement.

So here are the data from the 206 study, the dose-finding study, for the dose group that ultimately went on the Phase 3. So you can see that the exact categorical agreement was 87 percent, and the weighted kappa was .85 with 90 percent confidence intervals of .76 to .95. So that at the time seemed to look pretty good to us.

So in summary, for the entire Phase 2 program, not just the 206 study, the 1.5 microgram per kilogram IV bolus dose of binodenoson produced equivalent coronary hyperemia as intra-coronary adenosine.

There was strong image concordance with adenosine, as you saw in the 206 trial; lower prevalence and intensity of the common adenosine adverse effects, consistent with the A_{2A} selectivity in data I did not show you let from Phase 2. The effects were reversible with aminophylline, which is a competitive antagonist at the adenosine receptor. And aminophylline is commonly used in dipyridamole and sometimes in adenosine studies to turn off the effect, and that's important to know.

In the bronchospasm study, there was a decreased potential to induce bronchoconstriction in patients with mild asthma. Adenosine is contraindicated in patients with reactive airways disease, and the selectivity of this agent suggests thought it might be safe in those patients. And this was the first step in taking people with mild asthma and showing that there's no change in pulmonary function testing.

So that's where things stood at the end of Phase 2. And then we moved on to the Phase 3 pivotal program.

So the overall efficacy objective envisaged for the Phase 3 program was to demonstrate concordance between SPECT myocardial perfusion images acquired with binodenoson and SPECT myocardial perfusion images acquired with the active comparator, adenosine, as determined by independent, blinded expert readers.

The safety objectives were to evaluate and compare adverse effects, tolerability, side effects between binodenoson and adenosine, including the incidence of second or third degree AV block; the incidence and intensity of the commonly reported side effects from adenosine; and to get scoring or using tools to assess patient preference for one agent or the other, and how much the study bothered them, again, all with the idea that the selective nature of the A_{2A} adenosine receptor stimulation would reduce side effects compared to adenosine, and then of course to compare vital signs, ECGs, and all clinical laboratory and other general safety data.

Initially there were two identical studies, which we call Study 301 and 302. Both of these were multicenter, risk-stratified, randomized, double-blind, double-dummy, active-controlled, two-arm crossover designed studies. And as in the 206 trial, each patient completed two blinded pharmacologic stress SPECT perfusion imaging procedures in random order within one week.

The key inclusion criteria for both of these studies are shown here. These patients were clinically referred for an adenosine SPECT study on the basis of the history of chest pain. They were people who were on their way in to a nuclear cardiology laboratory for an adenosine SPECT study.

They had to be 30 years of age or older; some chest symptoms, typical or atypical angina; and importantly, we targeted populations across the spectrum, a pretest likelihood of coronary disease, to achieve what we thought would be a representative clinical population sample. And here are the targeted populations: 5 percent low likelihood, 45 percent intermediate likelihood, 25 percent high likelihood, and 25 percent known CAD.

This spread was as requested by or after consultation with FDA. And we were asked to enrich the population with intermediate likelihood patients, which makes sense because those are the patients who most benefit from noninvasive stress imaging tests. And these categories were based on American College of Cardiology/American Heart Association likelihood descriptions and categories.

The key exclusion criteria, among the many, are shown here: MI within 30 days; revascularization within three years unless there was new angina. Of course, if patients had a contraindication for adenosine reactive airways disease, they couldn't be in the trial because the patients were receiving adenosine. A severe LV dysfunction or advanced heart failure were also exclusion criteria.

The design, the general design, was fairly similar to what I showed you a few moments ago about Study 206. Eligible patients were randomized to a sequence, either adenosine first followed by binodenoson, or bino first followed by adenosine. But all patients received both studies.

Again, extensive efforts and training with the sites to create identical -- so that they would use, from one imaging procedure to the next, identical imaging protocols, cameras, isotopes, doses, et cetera, to minimize variability in the acquisition methodology and parameters.

Now, let me show what happens after the imaging studies were completed, sort of the tail end of the protocol.

So after the second imaging session was completed -- remember that these patients were referred for an adenosine SPECT study. So after the completion and all information data-gathering of the second procedure, the sequence was unblinded to the site because they needed to know which was the adenosine because they needed to read it and give the information to the referring clinician so they could make a management decision, because the sequence was unblinded at that point.

The adenosine data were given to the clinicians and the referring physician, of course, and medical management decisions, catheterization, no catheterization, et cetera, were based on the adenosine data that were ordered, plus all other clinical information.

All patients returned to the site for a follow-up visit one to four days later, at which time the questions regarding patient preference were done. The patients were still blinded at this point. And then the patients were followed out to 60 days, so at 30 days and at 60 days, to capture any information on clinically driven angiography that was done and any outcome events -- death, myocardial infarction, revascularization -- within those 60 days.

Now, during this time, the images, all of the images and any angiographic data that were available, were sent to core labs -- different core labs for the images, core labs for the angiograms --

for blinded analysis for, then, the data that we'll be showing you today.

Now, as I mentioned the drug administration was done in a double-blind, double-dummy way because so central to this was the demonstration of a reduction in side effects, it was really important that the drug administration be rigorously blinded. So this is an illustration of the double-blind, double-dummy drug administration.

At one of the imaging sessions, the patients received a placebo bolus, 30 seconds, followed by a six-minute infusion of adenosine at the FDA-approved dose. And this is the labeled administration of adenosine by the FDA labeling.

At the other imaging session, they received a binodenoson bolus and then a placebo infusion over six minutes. In both sessions, the radiopharmaceutical thallium, sestamibi, or tetrofosmin, was given at minute 3 after completion of the bolus because if it was adenosine active, that's the correct time to give the isotope; and if it's binodenoson, this is the correct time to give the isotope because that is clearly within the peak hyperemia that was seen in the prior Phase 2 studies. So this was very rigorously double-blinded and double-dummied.

Now, the image analysis in the entire Phase 3 program was done differently than I showed you for the 206 trial because this was done in complete compliance with FDA guidance for industry for image analysis in pivotal clinical trials.

The readers were independent. They read by themselves. They had no knowledge of other readers' interpretations. They were blinded to all treatment and patient data except for gender, age, and radiopharmaceutical. And the readings were done separated. In other words, each patient had two studies. One of those studies was read at one time point. The other study was not put into the reading queue until at least two weeks later for the reader to see it, again so they were completely separated in very randomized order.

Now, the images from the same patient were displayed on a monitor. The images from a patient were displayed on a monitor and scored on the electronic case report form. The quantitative program was available for the readers to look at, but the readers themselves were scoring the segments, and the electronic case report forms were completed on a separate monitor.

So unlike in the 206 trial where the readings were done side by side, the readers would read one study from a patient at one time point, and then separated by at least two weeks, they'd see -- it could be a month; it could be two months -- they'd see the other study from the same patient, and again, in complete compliance with FDA guidance.

Now, the hypothesis in Study 301, as you've heard, was we assumed that the true agreement of it was -- the metric for the statistical analysis was based on a weighted kappa analysis. We assumed that the true agreement for the weighted kappa statistic would be .75 point estimate.

The concordance between the binodenoson and the adenosine images would exist if the lower bound of the 95 percent confidence interval for the weighted kappa between the categorized SDS categories that I showed you before, generated by the blinded readers, was greater than or equal to .61.

Now, where did this come from, and this? It was based on the results of 206, as well as a review of the literature and some other analyses that we did. So this is where we started with Study 301.

Now, here are the population sample demographics, a good mix of genders, again just in Study 301. Age, 63; reasons for referral, mostly chest pain. A small percent of people had prior MI or revascularization.

On the bottom, these are our targeted populations across the likelihood categories. And on the right is the actual population broken down into those categories, which was pretty similar to the targeted populations. So a representative sample of patients coming to nuclear cardiology laboratories.

Now, here are the results in the 4x4 table. Again, adenosine across the top. These are the initial results that we saw, the primary efficacy result. No ischemia, mild, moderate, severe, by these

SDS score categories for adenosine across the top and binodenoson along the left column.

Now, much lower than we had anticipated, the weighted kappa was .24 with 95 percent confidence intervals, down to .14 and on the upper bound .34. The exact categorical agreement was 57 percent in this study.

Now, there are several things to note on this slide. First, that the categorical disagreement -- in other words, the patients who live below and above the diagonal, above where adenosine showed more ischemia and below where binodenoson showed more ischemia than adenosine -- the categorical agreement seems to be fairly evenly distributed, which means it is more or less equally likely that one agent or the other would show you a larger summed difference score. And this symmetry also suggests that there's not necessary bias here in this analysis.

Note also, and we'll talk more about this during Dr. LaVange's presentation, that the preponderance of patients live in this upper left-hand corner, normal or only mild ischemia, and a relatively smaller amount live down here in the lower right-hand corner.

Now, this is driven by the population. When you target intermediate likelihood and low likelihood people, they don't often have a lot of ischemia. But it is a representative sample of patients coming to a nuclear cardiology laboratory. So we'll have more to say about that point in a few minutes.

Can you go back one, please? Thanks.

When you have this type of disagreement, it's important potentially to have some kind of a gold standard; which one is right. And you can't necessarily assume that this is right and this is wrong or this is right and this is wrong.

So among the 300-plus patients who are enrolled into the 301 study, 50 of them, or about 15 percent, went on to angiography on the basis of their clinical data and the adenosine data, which of course was part of their clinical management and what they were initially referred for.

So here are the angiographic data from the 50, 5-0, patients in the 301 study who went on to angiography. Again, the angiographic data was analyzed in a core lab, the binodenoson and adenosine data analyzed in a core lab.

Here, normal and abnormal, abnormal refers to a greater than or equal to 50 percent stenosis on quantitative analysis in a blinded angiographic core lab. Normal and abnormal for the images mean summed difference score greater than or equal to 2, in other words, some degree of ischemia.

Sensitivity and specificity for binodenoson in this group of patients were 70 percent and 70 percent. Sensitivity and specificity for adenosine here in this group, same group going to angiography, 63 percent and 48 percent. And I'll note that these numbers are not too dissimilar from the labeled sensitivity and specificity for adenosine, which is 64 and 54 percent.

So this is what we had for angiography in the 301 data. But there was some signal reflecting the A_{2A} selectivity, as we had anticipated. These are the side effect data for binodenoson and adenosine. No heart block seen. Flushing, chest pain, dyspnea, all numerically reduced, as you would expect from the more selective A_{2A} agent. So there seemed to be a favorable signal in terms of side effects.

So the key findings from Study 301 at this point was that the prespecified kappa threshold was not achieved. However, there seemed to be compatible distribution above and below the diagonal, suggesting a similar degree of disagreement, as it were.

Sensitivity and specificity for angiography, and in data I didn't show you for a small number of clinical outcome endpoints followed out to 60 days, were comparable for binodenoson and adenosine, and the data suggested comparability between the images and when a gold standard was available. And certainly there was an improved side effect profile or tolerability profile achieved with binodenoson at this point compared to adenosine.

So at this point we had a kappa that didn't achieve the threshold; seemingly variability or symmetrical disagreement, as it were, and this caused us to go into a series of investigations to try and understand this and come up with some solution.

So with that, I will turn it over to Dr. LaVange to discuss the statistical considerations.

DR. LaVANGE: Thank you, Dr. Udelson, and thanks to the committee for allowing me to discuss the statistical considerations involved in developing the analysis strategy for Phase 3. I'd like to focus on three items in my presentation.

First, I will review the kappa statistic from Study 301, as well as results of the kappa analysis from an external study of a related compound, namely regadenoson.

Second, I will present the intra-class correlation coefficient. I will present this as the continuous data counterpart of the kappa statistic to assist in our understanding of what happened in Study 301, recognizing that the summed difference scores are essentially continuous. And then finally, I will provide the rationale for a change in the primary efficacy analysis of Studies 302 and 305 to a clinical equivalence analysis.

This slide represents the three-way concordance of the 50 subjects Dr. Udelson just mentioned for which we had results of myocardial perfusion and measures of ischemia available from binodenoson, adenosine, and angiography, all three measures.

The three-dimensional figure shows how well each method performed relative to angiography as well as each other. So here we have the binodenoson versus angiography results, with the sensitivity, specificity, and weighted kappa statistic; here, adenosine versus angiography, similar statistics; and then the two agents against each other.

The nodes on this cube represent three-way concordance where the preponderance of subjects are, all three abnormal, all three normal, as well as two-way concordance at the other nodes.

This figure shows that binodenoson appears to perform well relative to angiography, aside from how the kappa statistic reflects agreement between the two agents.

Shortly after Study 301 results were available for binodenoson, the efficacy results for another related compound, regadenoson, were available in the literature. Regadenoson has been recently approved for an indication similar to that targeted for binodenoson.

This table shows the published efficacy data on regadenoson. In this trial, which was part of the clinical development plan for regadenoson, patients were first to receive adenosine and then following by either adenosine or regadenoson in random assignment.

A criterion different from the kappa statistic was the basis of this trial's successful analysis. However, using the published data, we were able to construct a weighted kappa statistic that was similar to the kappa statistic for the Study 301 primary analysis.

Notice that the weighted kappa statistic for the adenosine/adenosine randomization group presented here had a moderate size of .48, and for adenosine/ regadenoson a moderate value of .50.

The upper bound of the confidence interval in both cases was less than the prespecified criteria for kappa in study 301, namely .61. In fact, the entire confidence interval is to the left of the criteria in both cases.

Now, this next display shows the 4x4 frequency table that was the basis for the kappa computation in Study 301, and there some dilemmas with the data structure presented here.

In computing the weighted kappa statistic, a patient is considered to be in full agreement if the categories assigned from the two methods are the same, and those patients would lie on the diagonal. Here are the four categories that the summed difference score was categorized into.

So, for example, a patient here is considered to be in full agreement. The binodenoson and adenosine summed difference scores were 8 and 5, respectively, for a difference of 3.

In contrast, this patient is considered to be in disagreement by one category because the agents classify the patient as moderate and mild, which is a different categorization. However, it happens in this example that the summed difference scores for binodenoson and adenosine are 4 and 5, differing in 1, which is less than the difference of the patient that's considered in full agreement.

So we believe that there is some loss of information in going from the summed difference score, which takes on discrete values -- in the Study 301 case, the values were from 0 to 20 -- and taking those discrete values and categorizing them into the four categories for purposes of computing the kappa statistic. This type of inconsistency illustrated here can negatively impact the utility of weighted kappa when it's used as a measure of concordance.

The weighted kappa statistic is statistically recognized as essentially the same as the intra-class correlation coefficient, which is the usual measure of agreement for continuous determination such as the summed difference score. Further understanding of the issues with the use of kappa can be gained by understanding the structure of the intra-class correlation coefficient, or the ICC, as shown here.

The ICC, as assessed with some different scores, has two components of variance. The first is the method-to-method or test-retest variance component, which is assessed within subjects, and it's denoted by sigma squared w in the numerator of the right-hand term.

The second component represents the heterogeneity of the population, or the patient-to-patient variance component, denoted by sigma squared s. And it's part of the total variance, which is in the numerator, and this ratio is subtracted by 1 to yield the intra-class correlation coefficient.

The ICC approaches 1 with better concordance, and it approaches 0 with less concordance, just as the weighted kappa statistic does.

Now, from the formula, it's clear that the reliability of a test will increase as the method-to-method variance component decreases. However, the extent of homogeneity in the population, which in our case is represented by a skewedness towards normal and mild cases, will limit the magnitude of the ICC even if the method-to-method variance is small because of the construct here.

So a better measure of the performance of a stress agent in this particular scenario would be a criterion that directly addresses the method-to-method variance component since it is a measure of agreement in its own right.

Such a criterion, based on the method-to-method variance component, can be specified in terms of a two-sided confidence interval about the mean of the paired differences between the two agents in the summed difference scores.

Now, here D corresponds to the difference for a single patient between the binodenoson and adenosine summed difference scores, and the mean, D bar, is the mean of these paired differences within-patient differences across the study population. This method-to-method variance component is represented here, and it governs the length of this confidence interval.

By requiring that this confidence interval lies wholly within an interval of minus delta and delta for some suitably small value of delta means that both the paired differences will be near each other. The mean paired differences will be zero. So the means of the two agents, adenosine and binodenoson, will be similar.

It also requires that the method-to-method variance, the variance component here, is small since that governs the width or the length of the confidence interval.

Now, the way in which a criterion based on a confidence interval works in terms of assessing agreement is illustrated on this slide. So these four figures illustrate the performance of the confidence interval criterion.

Assuming that a suitably small value for delta has been specified, a successful result is provided if the 95 percent confidence interval lies wholly within the interval minus delta and delta. And here are two examples where equivalence would be inferred. The two agents would be considered equivalent based on this confidence interval criterion.

When the confidence interval criterion is not met, then the confidence interval is exceeding the interval minus delta delta either on both ends or on one of the two ends. Either way, in this case the two agents would be considered to be not equivalent.

So equivalence here, based on this confidence interval criterion, means that the method-to-method variability is sufficiently small such that the two stress agents provide equivalent interpretations of their respective images. And note that the criterion for the confidence interval to be successful will only happen if the entire distribution of the paired differences are very tightly distributed about zero. And this we can see from the next graph.

So this is a distribution of the within-patient differences between binodenoson and adenosine with respect to the reader-generated summed difference scores from Study 301. The distribution of the paired differences here -- not the individual scores but the within-subject paired differences -- is centered near zero. And in fact, the mean paired difference is .15 based on the subjects in 301.

The tails of the distribution ramp off fairly quickly on both sides, and the majority of the patients fall within a fairly narrow interval about the mean. This is consistent with a small method-to-method variance component from our previous slide.

The other thing to note on this graph is that the distribution is symmetrically distributed about the mean, which is near zero. And that symmetry indicates that there's no tendency from one agent to have values that are different from the other agent in either direction.

Now, in order to apply a confidence interval-based criterion for establishing agreement through this test of equivalence, the margin, delta, needs to be specified in advance. And usually the margin or delta takes into account both clinical and statistical information.

So in terms of clinical information, the literature indicates that a difference greater than or equal to 5 percent in the amount of ischemic myocardium is associated with increased morbidity and mortality.

In addition, the literature shows that a difference of 3 summed different scores units represents altered perfusion in approximately 5 percent of the myocardium.

In addition, literature from four prognostic studies supports that a difference on the other side of 3, or in excess of 3, summed different scores would represent a clinically meaningful difference. Therefore, in selecting a margin of equivalence, you want to be substantially less than 3 summed different scores.

In terms of statistical information, the standard deviations for the different scores from Study 301 and 206 were examined, and they are presented here. Based on all patients in Study 301, the standard deviations for binodenoson and adenosine reads were 3.1 and 2.8, about the summed different scores. And for the more severe patients, both the mean summed difference score and the variability increases, as would be expected. The standard deviations are in the range of 4.7 and 4.9. From Phase 2, we have standard deviations of 3.2 and 3.3, consistent with the 301 data.

You would want your delta, your margin of equivalence, to be substantially less than a standard deviation of the summed difference scores.

In addition, we had the ability to look at rereads of a set of images from Study 301. These are images where the same reader read the image twice at two different points in time. And the absolute differences on these rereads range from .6 to 2.3. So whatever you choose for delta, your margin of equivalence, you would want that to lie somewhere in this range, which represents, in some sense, intra-reader variability.

So based on the clinical and statistical information that were available, it was determined that a value of delta of 1.5 summed difference score units would provide a sufficiently narrow interval for the test of equivalence between the two agents. This value is one-half of what was considered in the literature as the lower bound for a clinically meaningful difference, namely, 3 summed difference score units.

It is also approximately one-half of the standard deviation of the summed difference scores from Phase 2 and 3 studies. And finally, it falls within the range of intra-reader variability, as estimated by the rereads of a subset of 301 images.

Evaluating equivalence requires that the difference between the two methods have a confidence interval that lies within the bounds of minus delta delta, and that delta has to be prespecified in advance before you unmask and conduct your data analysis.

However, once you have conducted the data analysis, the extent to which the observed confidence interval may actually lie in an interval narrower or internal to minus delta delta would then provide evidence that the reliability between the two agents is even stronger.

So the primary efficacy analysis was revised to a clinical equivalence analysis based on the confidence interval criterion. The criterion directly addresses method-to-method variability, and success based on this criterion supports similarity of interpretation of the images for the two stress agents in a sense that is similar to that which applies to pharmacokinetic equivalent studies that are based on quantities such as area under the curve.

The revised primary analysis consists of two parts. First, the 95 percent confidence interval as just described, the mean paired difference and summed difference scores from binodenoson and adenosine must lie within the interval minus 1.5 and 1.5. This criterion ensures that the means are similar under the two agents and that the within-subject variance or method-to-method variance is sufficiently low.

The second component is that significantly fewer than 10 percent of the patients have extreme discordant results, where extreme discordance is defined as the two corners of the 4x4 table that the kappa statistic was based on.

This means that the upper bound of the 95 percent confidence interval, about the proportion of patients who have an abnormal read on one agent and a normal read on the other, severely abnormal and normal, has to be fewer than 10 percent. So that confidence interval has to exclude 10 percent, which means the actual percent of patients has to be much less than 10 percent. This component, the second component, guards against extreme differences cancelling each other out in the computation of the mean paired differences upon which the equivalence test is based.

The revised primary efficacy analysis was handled as follows. First, it was applied retrospectively to Study 301 data. This analysis is exploratory because the study had already been unmasked. The preplanned primary analysis, based on the weighted kappa, had failed for reasons we believe are related to the limitations of the kappa statistic in this scenario, as previously described.

Second, the revised analysis was invoked prospectively for Studies 302 and 305 by protocol amendment, and that protocol was put into place while those studies were still masked.

The original primary analysis, based on the weighted kappa, was retained for completeness, and the original analysis objective of showing concordance between the two imaging agents remains unchanged with this strategy.

Operationally, the analysis plan for Study

302, which had already been written, was revised. The protocol was amended to reflect this analysis strategy change. Study 305 had not yet had the analysis plan written, so the analysis plan was prepared to reflect the clinical equivalence analysis as primary, and the protocol was amended.

Both studies remained unmasked, and in fact the images had not even been merged from the core lab to the clinical database when this took place.

At this time I'd like to turn the presentation back to Dr. Udelson, and he will give you the results of the remaining Phase 3 studies.

DR. UDELSON: Thank you, Dr. LaVange.

I just wanted to start with the timeline that Dr. Carter showed you earlier in the presentation, just to emphasize the final points that Dr. LaVange made, that the revised analysis plan and the protocol amendments were put into place for Studies 302 and 305 prior to database lock and unblinding in 302, and in fact prior to the images being read in Study 305.

In our view, what we were changing was the analytic methodology of the data. Nothing else about the trials changed, the inclusion/exclusion criteria, the population samples, the image acquisition, the image analysis, et cetera, and the side effect analysis, of course.

So the design of Study 302 was exactly the same as Study 301 that I showed you before. Patients were randomized to a sequence. All patients received both a binodenoson and adenosine study within a week. After the second study was completed and data were acquired, the sequence was unblinded to the site again so that the adenosine study could be read clinically, as that was ordered for the patient.

The medical management was based on the adenosine data. Follow-up one to four days later. And then 30- and 60-day follow-up to capture angiographic data and any outcome events. Again, in this trial, all of the images and the angiographic data, when available, were sent to core labs for blinded analysis.

Now, the 305 study incorporated a different feature up front. The patients were randomized to a sequence, as before, but in a 3:3:2 ratio.

Some patients were randomized to an adenosine/adenosine arm, so each patient had adenosine twice within a week to assess the test/retest variability and create context for the adenosine/ binodenoson comparison. All of the other features, the image acquisition, core lab analysis, side effects analysis, et cetera, were exactly the same as in the 302 study.

This slide demonstrates the population sample demographics in Study 302 on your left and Study 305 on your right. Again, good mix of genders. Age typical for patients seen in such a lab. Most patients referred for chest pain. And at the bottom, here are targeted pretest likelihood categories on the left column, and on the right column near it are the actual percent of patients in those categories within the study.

You'll note that the proportions here in Study 305 are slightly different, and this was based on an observational outcome study in 5,000 patients that we had performed between here and here with general pharmacologic stress testing to reflect international populations, and the actual patients in the trials shown here. And, again, in all of the trials, a large number of patients had an intermediate pretest likelihood of coronary disease because, again, those are patients who benefit from noninvasive stress testing.

Now, here are the raw data, as it were, the SDS difference, as Dr. LaVange explained, in the 302 study. So the X axis is the binodenoson summed difference score minus the adenosine summed difference score. The Y axis is the number of patients. And as you can see, as she showed you in the 301 study, most of the patients are clustered within the small numbers and rapid tail-off, kind of a symmetrical distribution about the large number of patients. And just an illustration, a zero difference, that might mean a patient who had two normal scans, you know, a 0 SDS and a 0 SDS; or it might be a patient who an 8 or an 8, or a 12 and a 12. So this is the binodenoson minus adenosine SDS difference in Study 302. The mean was minus 0.09.

Here are the data in Study 305, now almost 400 patients. Again, the X axis is binodenoson minus adenosine. A large number of patients clustered within the small numbers, tailing off somewhat symmetrically. The mean difference is minus 0.68.

Then finally, in Study 305, these are the patients who had adenosine twice, so adenosine for the first study, adenosine second, in a double-blind, double-dummy manner. Now on the X axis is the adenosine minus adenosine SDS difference. And like the others, you see a cluster of patients around the small numbers. But you see tails in both directions.

So when you do the same study twice in a week in a highly controlled clinical trial environment where the images are read under sort of a regulatory reading environment, this is what you see. There are some patients who have extreme differences in one direction. Some patients, small numbers, have extreme differences in the other direction. The mean SDS difference here was minus 0.12.

So essentially, these are the histograms and the raw data from which the final primary endpoint analysis of the SDS difference, using the clinical equivalence criteria and the confidence interval margins that Dr. LaVange discussed, was constructed.

So here are the bounds that were discussed for the 95 percent that the confidence intervals must fall within. Here's from Study 302, the point estimate which was on the earlier slide, minus 0.09, and the confidence intervals fall well within the bounds of the minus 1.5 to 1.5 SDS units.

Here are the data from Study 305, the binodenoson/adenosine comparison. Again, the data from the previous slide, now with the confidence bounds falling well within the minus 1.5 to 1.5 equivalence margins, and then the adenosine/adenosine data in Study 305, shown here, the point estimate and the confidence intervals again falling well within those boundaries. So this was one component of the revised primary efficacy analysis for Studies 302 and 305.

Now, as Dr. LaVange mentioned, the other component required that fewer than 10 percent, or the upper bound of the confidence interval, was less than

10 percent of the number of patients who fell into the extreme difference categories to guard against too many patients with extreme differences cancelling out, creating a mean difference of zero. And I think from the histograms you could see that the number of those patients were relatively small. But here are the numbers.

So in 302 the patients in the extreme off-diagonal cells were 3 percent, only 11 of the 374. We'll talk about the other data in a few slides from now. In the 305 study, there were 12 total patients, or 3 percent of the population. And in the adenosine/adenosine comparison, there were 5 patients out of the 138 that were randomized to that sequence, where 4 percent of patients fell in the extreme corner. So in all three of the comparisons, there were well less -- at least the point estimate was well less than 10 percent that had been part of the hypothesis.

So in summary, for the revised primary efficacy analysis in Studies 302 and 305, we believe these data demonstrate concordance between binodenoson and adenosine pharmacologic SPECT images from myocardial perfusion imaging procedures based on leader-generated summed difference scores because the 95 percent confidence intervals around the mean paired difference were well within the prespecified equivalence margins of plus or minus 1.5 SDS units. And well less than 10 percent of patients had extremely discordant results, that is, those at the extreme corners of the 4x4 cross-tabulation tables.

Now, those patients with the extreme differences are of particular interest and it is interesting to know which was right, as far as you could know that. So how many of those patients had some independent gold standard?

So here are the data that I showed you before. Four percent and 3 percent of patients in the Studies 302 and 305 fell into those corners. When you look back at the regadenoson data, which Dr. LaVange showed you, the data were fairly similar.

When they did adenosine/adenosine twice, 6 percent of those patients fell into extremely discordant results, 4 percent in the regadenoson/adenosine comparison. So these numbers show up consistently when you read images in a regulatory environment by FDA guidelines.

Now, in the 302 and the 305 study, of the 22 patients -- across the 301, 302, and 305 studies, of the 22 patients who fell into the extreme corner where adenosine was severely ischemic and binodenoson was normal, 8 of those patients went on to clinically indicated angiography. And remember, the angiography was based on the site reading, not the core lab reading.

So 8 of those patients who had a severely ischemic adenosine scan, normal binodenoson scan; and of those 8 at angiography, 4 were normal, in other words, the binodenoson was correct and the adenosine was wrong; and 4 were abnormal, in other words, the adenosine was correct and the binodenoson was incorrect. So half and half.

Now, in the next few slides, I'd like to display much of the other data on the levels of agreement and the kappa statistic in the three studies.

In this slide, what is here, these are the reader-generated scores, the summed difference score that we've been talking about, as well as the summed stress score, which, as I mentioned earlier, is the most powerful prognostic predictor, actually, in observational trials, and the summed rest score. So these are the reader-generated data. And now for the first time I'll also show you the computer-generated data. So no human eyeballs, just computer-generated data all on its own.

So for the reader, the summed difference score is shown here. This is the binodenoson -- or this is the difference between the SDS scores of the two comparisons. These are the equivalence margins of minus 1.5 to plus 1.5, as we mentioned before. Here are the reader data that I essentially showed you already before.

Now, the summed stress score looks very similar. All of the data, the confidence intervals widely overlap. The resting scores are interesting because, you know, when you do two rest studies separated by a week with no stress intervention, in some ways this is the limits of the agreement that you could ever see in this kind of reading environment. And you can see there's some confidence intervals around this; all the data line up.

Now, the computer reads, summed difference score, again, lots of overlap between the data from the trials and the adenosine/adenosine data. And the summed stress score is read by the computer alone; also, overlap between the confidence intervals from Studies 301, 302, 305, and the adenosine data.

Now, as discussed earlier, in discussions with the FDA we had also said that all of the kappa data would be displayed and computed for all of these trials.

So here now on the X axis is the weighted kappa statistic value. On the Y or in the column here are the different readings that I mentioned on the previous slide -- the reader-generated summed difference, summed stress, and summed rest score, and just the computer-generated -- no human -- summed difference, summed stress, and summed rest score.

Here are the weighted kappa data on Study 301, which I showed you originally. And here is Study

302 and 305. Here's the adenosine data. Note the summed stress score up here; almost complete overlap of the confidence intervals, all of them falling in the point estimates, .5 to .6 range. And again, the resting scores, again, provide some context because this is really the limit of how good kappa can be when readers in that kind of reading environment read rest images where there's no intervention, no stress intervention. So again, .5, .6 range.

The computer data looked fairly similar. Summed stress scores are moving up a little bit along the kappa scale, weighted kappa scale, but overlap in the confidence intervals here.

Now, in this slide we display the absolute paired difference. Now, previously we showed you histograms that showed both negative differences and positive differences.

Here's the absolute paired difference in the reader-generated summed difference score to perhaps get a better example of some of these differences. And as you can see, the majority of patients really cluster around differences between the two agents of

0, 1, and 2, tailing off toward the larger scores and then going out to 10 to 20, et cetera, just listed as greater than or equal to 10.

Now, the context is the blue, which is the adenosine/adenosine comparison. And just sort of qualitatively, you can see that throughout the distribution of differences, as it were here, that adenosine/adenosine comparison is little different from the binodenoson/adenosine comparisons.

Now, this slide is a copy, essentially, of Figure 2 in the FDA briefing document, not the sponsor briefing document, but the FDA materials that you received. And it's a cumulative distribution function of the Study 305 binodenoson/adenosine difference scores. So the X axis here is the spread of the absolute difference scores, the difference between binodenoson and adenosine summed difference scores. And this is the Y axis, just a cumulative distribution.

Now, in the text, it was correctly noted that, you know, we said that a summed difference score of more than 3 was clinically meaningful, as

Dr. LaVange told you, and that if you follow this up here, about 25 percent of the population in this study fell above a level of 3, a difference of 3 in the summed difference scores, suggesting susceptibility to a difference in diagnosis, which is completely correct, of course.

So I'd like to take the liberty of adding in the adenosine/adenosine data here in the same format, the cumulative distribution function across the differences, which is shown here. And if you do some analysis on the difference between these two curves, this represents the 95 percent confidence interval of the difference in the means of the absolute differences, so a difference of a difference of a difference. And as you can see, the 95 percent confidence intervals are very narrow, do not include 1, and include 0, suggesting that these curves, the absolute differences, the distribution of absolute differences, is similar between binodenoson and adenosine and doing an adenosine study twice.

Now, to make this a little more complicated, I'll also add in Study 302 and 301 cumulative distribution function. And without doing that same analysis but telling you that it's about the same if we have done it, that the curves, all the cumulative distribution, the differences look very similar across the studies and very similar to doing an adenosine study twice.

Now, let me go on to talk about the angiographic data because much of the discussion today involves differences between two agents. And, you know, if one is showing more abnormality than another, it might be inferred that one is better than another. But in essence, is there an independent gold standard? And there was in a subset of the patients, the angiographic data.

So as I described, and as is clear in the documents, patients were referred to angiography on clinical grounds. It was not protocol specified because that was not the primary purpose of the protocol.

Based on clinical data and the adenosine SPECT data, these patients came into the protocols having been clinically referred for an adenosine SPECT study. So that did indeed drive the decision to angiography.

The angiographic data were analyzed in a blinded core laboratory. The binodenoson could be seen at the sites, but the sites were, of course, instructed not to use that data to drive any decisions because it's an investigational agent. And I'm sure many people did not because that's not appropriate or ethical.

Here are the results of the measures of accuracy, sensitivity, specificity, et cetera, among the 204 patients across the Phase 3 trials that underwent angiography on clinical grounds. And this represents 15 percent of the entire population enrolled into this study. And that 15 percent is very typical of a practice in nuclear cardiology, and you can pull that same number out of large databases. So 15 percent of the patients went on to angiography.

The columns are -- this is the reader-determined summed difference score and the computer-determined summed difference score for binodenoson and adenosine. A positive study by SDS was a score greater than or equal to 2, in other words, ischemia, and a positive angiography was a 50 percent or more stenosis. Or if the core lab happened to think the stenosis was less than 50 percent, if the patient was revascularized clinically, we called that positive as well.

So as you can see, the point estimates, the sensitivity is a little bit higher with adenosine in the studies, but the specificity is lower. The point estimate, positive predictive value is somewhat similar, negative predictive value is somewhat similar, and overall accuracy across the trials somewhat similar.

Now, again, it's instructive to look at the patients who had an imaging disagreement. And what did the independent gold standard have to say about that?

So this slide represents data among patients who went to angiography who had a disagreement in binodenoson versus adenosine summed difference scores. So on the top, if this disagreement happened by the reader-determined summed difference score and the patient went on to angiography, 56 percent of the time the binodenoson study was correct, based on the angiography, and 44 percent of the time the adenosine study was correct.

If the disagreement on SDS was by just the computer program read, 48 percent of the time binodenoson was correct and 52 percent of the time adenosine was correct. So I think it would be fair to say that when disagreements happened among those patients who went to angiography, one agent was right about half the time, the other agent was right the other half the time.

Now, we also analyzed these data by ROC curve areas. And here the binodenoson data are shown in the green, the adenosine data shown in the gold. And this is for the 204 patients who underwent angiography. And so the area under the curve represents the discriminate ability of the imaging to discriminate the presence from the absence of a 50 percent or greater stenosis. And you can see that the curves are fairly similar. Confidence intervals overlap. If anything, the binodenoson is a little bit higher.

On the next slide, this is the analysis using -- this is an ROC analysis for what we call the clinical endpoint in all patients across the Phase 3 trials. And what this means, positive is clinically driven revascularization, myocardial infarction, or death.

Now, there were very few myocardial infarctions and there were no deaths, so this is predominately clinically driven revascularization, which of course in part was driven by the adenosine data but not by the binodenoson data. And as you can see, the curves are essentially superimposed to discriminate the presence from the absence of a clinical endpoint, predominately revascularization in these thousand-plus patients.

So the conclusions regarding efficacy, when you look at the totality of the data, the revised primary endpoint, and all of the other data, including the angiographic data, we believe that the binodenoson SPECT images provide comparable clinical information on the extent and severity of ischemia as the adenosine SPECT images. The degree of equivalence between the agents seems to be comparable, or is comparable, to that of performing adenosine SPECT imaging twice.

Those conclusions on efficacy we believe are supported by the totality of the data by both the reader- and the computer-generated summed difference scores and summed stress scores. There's clinical equivalence within the margins that we discussed, with a small number in the extreme corners.

We examined the absolute paired differences, as I showed you, the weighted kappa values, which we demonstrated, and importantly, the measures of accuracy for angiography in clinical endpoints were similar between the agents.

As I mentioned, the reader- and the computer-generated summed rest scores in the adenosine data both served for context and for reference to really the upper limits of agreement that were possible using these different analytic methodologies.

I'd like to move now in the last few minutes of my presentation to the safety and tolerability assessment. And again, remember that the reason for developing these agents is the selectivity at the A_{2A} receptor and the potential to reduce the bothersome side effects of pharmacologic stress testing.

So the safety objectives of the Phase 3 program were to evaluate the incidence of second or third degree AV block; compare adverse events between binodenoson and adenosine, particularly side effects and tolerability; evaluate the patient bother, how much the study bothered the patient and which agent they had a preference for; and of course, to compare vital signs, ECG changes, and clinical laboratory data.

There were tools used to assess some of these parameters: a visual analog scale, a 10-point validated visual analog scale adapted from McGill, applied to the intensity of the common side effects: flushing, chest pain, dyspnea, nausea, headache, abdominal discomfort, and dizziness.

On the patient assessment of bother and preference while blinded, the bother question, how much did this study bother you, was asked to the patient following each myocardial perfusion imaging study; and the preference, which study did you prefer, while still blinded, was asked at the end of the second or at the follow-up after both imaging procedures had been completed.

Now, it's important to note that the tools that were used and the scales that were used and the questions that were used were validated by two independent validation studies conducted in patients undergoing adenosine pharmacologic stress imaging, not within these protocols but independently from these protocols. And the VAS tool was shown to be -- the response of it was shown to be valid, reliable, and responsive, and the bother measure was shown to be reliable and valid. And these data were published earlier this year.

Now, because we are interested in many different side effects, it was important that this was done in a rigorous way, accounting for multiplicity. So the order of analysis of the safety endpoints of interest were pre-specified for sequential testing to account for multiplicity. And the order, the prespecified sequence, was second or third degree AV block, the bother and the preference question, and then the incidence and patient-rated intensity of flushing, chest pain, dyspnea, nausea, headache, abdominal discomfort, and dizziness, the common side effects from the literature and that had been seen in earlier trials.

Now, there was a statistical comparison of each one of these in sequence, and when a comparison in that sequence did not reach significance, no further inferential testing was performed, but the data are reported.

Here are the overall sort of general safety and adverse event summary from the over 1,000 patients in the three Phase 3 trials getting binodenoson and getting adenosine. Ninety percent of patients reported any treatment-emergent adverse events with binodenoson, 96 percent with the adenosine. The relation to study drug was similar.

The intensity, if you can see here, by the proportions, was shifted somewhat toward the more mild, with binodenoson compared to adenosine. There were no deaths across the Phase 3 program.

Serious adverse events were rare, less than 1 percent, 6 patients in each group. And treatment-emergent adverse events leading to study drug or study discontinuation were also infrequent, 8 patients with binodenoson, 11 patients with adenosine.

So here are the data from within the sequential testing analysis. From Study 302 and Study 305, binodenoson and adenosine, in the sequence of order of testing of the common or the usual side effects seen with adenosine testing, there was no second- or third-degree AV block observed with binodenoson. In fact, it's never been observed throughout the entire program; infrequent with adenosine, 3 percent and 1 percent.

Within both studies, the incidence of flushing was less. Chest pain was less. Dyspnea was less. And in Study 305, nausea was significantly reduced, but not in Study 302, although it was numerically reduced.

So after the final statistical significance was reached here, no further inferential testing was done in the remaining sequence. You do see here that headache was more common with binodenoson compared to adenosine in both studies.

Now, in this slide, that data on the previous slide was the incidence. This is the intensity of the side effects as rated by the patients while blinded, using the visual analog scale tool in study 302 and 305.

What you can see here, in this analysis when a side effect did not occur, a score of 0 was imputed, but the general pattern is the same if you take those zeroes out as well.

So here the intensity of flushing, chest pain, and shortness of breath was reduced significantly in both studies, nausea reduced in the 305 study, not quite significant in the 302 study. So, again, after these points, no further statistical testing was done. Note, however, for headache, the intensity when it happened seemed to be similar between the two agents.

Now, we also, as I mentioned, asked the patients which study did you prefer, study number 1 or study number 2, while the patients were still blinded. And in green is the binodenoson. Gold is the adenosine. Blue is no preference. And in both of the studies, about 70 percent of the patients preferred the study that turned out to be binodenoson, only 20 percent preferred adenosine, and about 10 percent had no preference, and the P values represent the difference in proportions.

In terms of the question, how much were you bothered by this test, the patients were asked to rate that answer on the scale of not at all, a little, some, or a lot. And here the binodenoson data are in green, the adenosine in gold. And as you can see in both Study 302 and in 305, there's a shift toward the "not at all" or "a little" with binodenoson, and a shift toward the "some" and particularly "a lot" with the adenosine data. And the differences in proportions was highly statistically significant in both of these trials. And you can see in the two different trials the patterns, in fact, were quite similar, as were the actual numbers.

Vital signs and EKG changes I'll show you in the last few slides. Mean changes in vital signs are shown here. A change in systolic blood pressure through 60 minutes were similar with binodenoson and adenosine, a drop of about 14 millimeters; diastolic blood pressure, similar between the two agents; heart rate increase, similar between the two agents. And of course, some of you are familiar with adenosine testing and performance, and this is generally what you see, of course, when you do an adenosine test.

Clinically important categorical changes in vital signs shown here are changes in blood pressure to less than 80, or a greater than 30 millimeter drop, similar between the two agents. Drop in diastolic pressure to those levels, similar.

No patient had bradycardia to less than 30 beats a minute, and tachycardia to an increased heart rate greater than 120, a little bit more often with binodenoson compared to adenosine.

Changes in electrocardiographic parameters shown on this slide. The heart rate data I showed you on the previous slide. Changes in the PR interval down a little bit with binodenoson, up a little bit with adenosine, consistent with its effects.

QRS interval, little change, no difference between the two. A QT interval by Fridericia's calculation, change from baseline plus 12 with binodenoson, plus 16 with adenosine, and no real difference between the two agents.

So the conclusions regarding safety and tolerability across these Phase 3 trials was that compared to adenosine, the more selective adenosine A_{2A} receptor agonist, binodenoson, demonstrated no second- or third-degree AV block that was observed.

Patient preference for binodenoson, less patient bother with binodenoson. So overall, I guess you could say a better patient experience. And then a significant reduction in the incidence and severity of flushing, chest pain, and dyspnea, the three common side effects seen with pharmacologic stress testing.

So I'd like to turn it back to Dr. Carter to summarize the benefits and risks.

DR. CARTER: Thank you, Dr. Udelson.

Mr. Chairman, my concluding remarks will keep us well within our timeline margins. I just want to assure you of that.

So ladies and gentlemen, the FDA has expressed concerns about the validity of the primary efficacy endpoint since we amended the statistical analysis plan of the pivotal studies during the conduct of our trials.

I believe that we've described how important learnings from the results of our first large clinical trial and newly available public data from other imaging studies justified the amendment on a sound and rational basis.

We've also shown that we applied the amended statistical methodology prospectively in the two primary efficacy studies, Study 302 and Study 305, while still blinded, and thereby appropriately preserved the validity of the integrity of the data used to define the efficacy profile of binodenoson. And for completeness, we've presented both the original as well as the amended analysis.

In order to compare favorably with a reference agent, binodenoson had to produce similar hyperemia to that produced by adenosine, and

Dr. Udelson showed you those data. He also showed that there were no clinical meaningful inconsistencies in the primary endpoints between Study 302 and the confirmatory Study 305.

Concordance between binodenoson and adenosine was prospectively demonstrated through an equivalence analysis in the same patients. And in this clinical setting, there was high agreement between binodenoson and adenosine in the assessment of the presence, extent, and severity of reversible perfusion defects; in other words, on the extent of inducible ischemia in patients with varying pretest likelihood of coronary artery disease.

When we examined imaging results against angiography, the true standard, the measures of accuracy for binodenoson were comparable to the same measures for adenosine.

We do believe that the totality of the data that we've presented today using multiple prospectively defined methodologies to demonstrate concordance provides compelling evidence that binodenoson and adenosine have similar clinical utility as pharmacologic stress agents for myocardial perfusion imaging studies.

We can conclude that binodenoson provides equivalent diagnostic information to adenosine, widely regarded as the best available pharmacologic stress agent in the U.S. And finally, the conduct of pharmacologic stress testing is simplified greatly by using a bolus dosing regimen as opposed to a six-minute infusion.

Moving on to the safety profile, we've shown that because of its selective pharmacology, binodenoson is associated with fewer and less severe subjective adverse effects than adenosine, and importantly, no reports of second- or third-degree AV block. In addition, the incidence and severity of adverse events of special interest is reduced.

So we've met a key objective of the development program, which was to demonstrate an improved safety profile over adenosine. Consistent with this, binodenoson was preferred amongst patients compared to adenosine, and on the whole, these patients tolerated binodenoson very well.

In conclusion, therefore, we believe that we've demonstrated that binodenoson fills an unmet need for a selective adenosine receptor agonist for use as a pharmacologic stress agent. This slide summarizes the key parameters of benefits and risks that were presented today. The bullets represent a favorable parameter.

Thus, binodenoson provides equivalent pharmacologic response and diagnostic information to adenosine, as shown up here. At the same time, based on events of special interest, the safety and tolerability profile is improved, and the conduct of stress testing may well be simplified. Of course, as mentioned by Dr. Udelson, we did see a numerical increase in headaches reported after binodenoson.

Overall, then, we believe that binodenoson has a more favorable benefit-to-risk profile than adenosine.

For the question and answer session, we're joined today by the following experts:

Dr. Rich Barrett was the project leader from the outset. He's now a consultant for King, and he'll help address your questions.

Dr. Edward Ficaro, who's president of INVIA Medical Imaging Solutions and an expert in computer analysis of SPECT images is also here.

And I'm delighted that Dr. Gary Koch, professor of biostatistics at the University of North Carolina at Chapel Hill and a colleague of Dr. LaVange, is also with us. Dr. Koch has been involved with this project for some time, and as you may know, he is an expert in the statistical treatment of observer agreement. And in addition, we can call upon other members of the King development team as needed. Thank you.

DR. HARRINGTON: Thank you, Dr. Carter, on a very thorough and on-time set of presentations.

We now will take a break till 10:15. We have a big panel, so I'd like people to be ready to go right at 10:15. We'll then have a half hour of being able to ask questions of the sponsor. Also, if panel members don't think that's enough time, I think we'll have plenty of time this afternoon as well to come back to the sponsor.

I'm now required to read this statement that we will take a short break. Committee members, please remember that there will be no discussion of the meeting topic during the breaks amongst yourselves or with any member of the audience. And we'd like to resume exactly at 10:15. Thank you.

(Whereupon, a recess was taken from 10:00 a.m. to 10:15 a.m.)

DR. HARRINGTON: All right. Now that Elaine's back, I'm confident we can keep up with the questions.

So we now have approximately a half hour for the panel to ask questions of the sponsor. And following that, we'll have a presentation by the FDA. Also, we'll have a half hour in which to ask questions before we break for lunch.

So I'd like to open it up to the panel. And if you could raise your hand so that Elaine and I can keep track of it.

Yes, go ahead.

DR. CARTER: I'm going to try to coordinate the Q&A from our side --

DR. HARRINGTON: Perfect.

DR. CARTER: -- to make it easy for us to get to answer your questions as succinctly as possible.

DR. HARRINGTON: Perfect.

DR. CARTER: So if the members of the committee could direct their questions to me, I'll find the right person to provide the answer.

DR. HARRINGTON: Great. Thank you.

Sanjay, why don't we start with you.

DR. KAUL: Well, thank you. I am sympathetic to your predicament where overestimation of agreement based on Study 206 led to setting a high kappa bar. And one reason might be related to paired assessment of images which, as you acknowledge, tends to inflate agreement.

Another could be related to the types of patients that were studied in the Study 206. I noticed that nearly two-third of the scans were normal if you go to slide CC-31. And even in the advanced MPI program with the regadenoson, the degree of agreement was higher for normal scans, about 84 percent, compared to average agreement of 62 percent. So had you chosen a more representative sample, you would likely have not set yourselves such a high bar to overcome.

So in that spirit, I'm trying to understand the rationale, both clinical as well as statistical, for the equivalence margin that you chose. Let me first focus on the clinical rationale.

Are you trying to suggest to me that a 5 percent or a 6 percent myocardial perfusion defect in a young nondiabetic male with a normal LV systolic function portends a higher risk than a 4 percent myocardial perfusion defect in an elderly diabetic individual with an EF of about 45 percent? That's the clinical part of the rationale.

DR. CARTER: So let's see if we can answer that first.

Dr. Udelson, would you like to take this question, please?

DR. UDELSON: Well, no. I think in populations that you would look at, in the data from some of the references that we showed, particularly the Cedars Sinai database where you work, more than 5 percent ischemic myocardium was associated with an identifiable increase in a large population of risk of death over follow-up, from 0 to 5 versus above 5. So that was one of the data.

Now, I completely acknowledge your point that the risk stratification is strongly influenced by the pretest probability, and that the same degree of abnormality in an elderly diabetic woman is associated with a certain risk, whereas in a young nondiabetic male with the same degree of myocardial ischemia, that's a different risk, a lower risk, because the pretest abnormality, you know, as you have written about eloquently, really drives that.

Nonetheless, we needed to come up with a rationale that was rigorous, based on data, and that could be applied to populations for, when you're really examining a continuous scale, what constituted reasonable agreement between two studies. And in looking through the literature, three SDS units, which translates into about 5 percent of the myocardium, based on multiple studies, based on the categories that have been used in many of these studies where a jump of 3 will almost always get you into another category, and based on the fact that if you use such categories in large populations, you see an incremental risk, that did seem reasonable.

You know, in practice, I think all of us -- and, you know, I teach this every day, the image adds to the pretest likelihood to create some post-test likelihood of disease or post-test likelihood of risk. So, I completely acknowledge your point.

So from a clinical perspective, the three SDS units came from multiple population-based studies.

DR. KAUL: In your data set, how many were diabetic? How many had an LV of greater than 35 percent but less than 40 to 45 percent? I'm trying to see if we can apply this clinically relevant difference within your data set.

DR. UDELSON: Hang on a moment. We'll see if we have the diabetics for you.

Okay. Obviously, we must have those data in the tables. But if we can get that for you and perhaps show it to you after lunch, would that be acceptable?

DR. KAUL: That would be fine.

May I ask about the statistical reasoning?

If I understand correctly, you chose 50 percent of the standard deviation based on some pilot studies.

What was the standard deviation of the SDS within the trials, the 301, 302, and 305?

DR. CARTER: Dr. LaVange, please.

DR. LaVANGE: In the 301 and 206 trials, the standard deviations of SDS range from about 2.8 to 3.3. If you want to put this slide up, this has the individual studies and each of the scans, and the standard deviation is in parentheses after the mean.

So 301, I guess, on the far right, is the information we had available when we were picking the threshold of 1.5 summed difference score units. And you can see the range there with the reader-generated scores for binodenoson and adenosine on the -- not the computer but the top two rows. 3.09 and 2.8 is what we were working on. And then we looked back at Phase 2, the 206 study, and the range was about 3.1 to 3.2

there as well.

DR. KAUL: So the question I had is why did you choose 50 percent of that? Is that an arbitrary cutoff, or is it based on a precedent? Why not just 25 percent?

DR. LaVANGE: I think you want to choose something that's substantially less than the standard deviation to indicate agreement as opposed to, you know, variability. The one-half was somewhat arbitrary. We could have chosen a third, and a third would have been about one unit. And in fact, when the other two studies were amassed, the interval of the 302 study did lie within 1 and 1. But the one-half itself was somewhat arbitrary, yes.

DR. KAUL: To put some clinical context behind that, can you tell me what odds ratio does it approximate? Half, .5 of the standard deviation?

DR. LaVANGE: I'll ask Dr. Koch to answer that.

DR. KOCH: Yes. Gary Koch, biostatistics department, University of North Carolina.

As far as I know, there was not an odds

ratio calculation to motivate that. But as was indicated in Dr. LaVange's presentation, there were two or three different arguments that was supporting the 3 as a target difference that had clinical meaning. And one wants to have equivalence margins that are less than half of whatever would be arguably something that had a clinical interpretation. And so that was where the one and a half came from. And as the data showed, it actually met one-third of that.

The reason why you choose a half is because you want to meet something that is closer to the null than closer to the threshold. So 3 is the threshold. If you meet something that is less than half of that, you're going to be meeting something that's closer to the null than to the threshold.

DR. HARRINGTON: Dr. Neaton?

DR. NEATON: I have a couple questions just to follow up on this one. Maybe we could ask the sponsor to come back to this issue because on page 51 of the report, there is an attempt to put this into clinical context, and I could not follow the arithmetic there. I think some of the numbers may not be correct.

But using the data from the Journal of American Cardiology paper in 2005, can you put into context kind of what a difference along the lines that you specified was in terms of relationship with future cardiac events, which I think you should be able to do from that paper? And there was at least an attempt to do that in your writeup.

I have a couple just very simple questions to make certain that we're all comfortable with the designs. And so was the allocation to the two sequences equal in the three studies, the AB and BA allocations?

DR. CARTER: I believe so, yes.

DR. NEATON: And were the comparability of the patients assigned to the two sequences equal? Can you kind of put something up to give us some comfort that the randomization was carried out and the integrity of it?

DR. CARTER: Let's see if we can get those data.

DR. NEATON: And then typically in a trial, a crossover study like this, one would like to see some measure of whether there was any kind of treatment by period interaction.

Did you look for that?

DR. CARTER: Whilst we get to that third question, perhaps I'll ask Dr. Udelson to come and talk about the demographics here.

Jim?

DR. NEATON: I'm not talking about the demographics. I'm just thinking about, you randomized people to two different sequences; were they comparable? And is there any evidence of an order effect where the pair difference is similar for those given adenosine first versus second? You know, just so we're -- I mean, I assume that was looked at, given the study design. But I didn't see it anywhere in any of the writeup.

DR. UDELSON: Can we have this slide up, please?

So these are the data from Study 305 where the patients were randomized 3:3:2 to the first binodenoson/adenosine sequence, adenosine/binodenoson, and then adenosine/adenosine. And these are the differences in the demographics across the three sequences in the first three columns.

DR. NEATON: All right. And the test for interaction for the primary outcomes that you looked at for your kind of stress score.

Were the stress score differences similar?

DR. KOCH: My understanding is that the sponsor did indeed fit traditional models to the crossover study with assessments of carryover effect, which is treatment by period interaction, and did not find any. But we do not seem to have a slide to show that.

Is that correct?

DR. NEATON: Maybe you could just verify that kind of during the lunch, too.

I didn't understand how the different readers were used, so that in terms of getting a stress score and the rest score and then the differences, were those scores averaged over readers or was there a single reader for each patient?

DR. CARTER: Jim, please?

DR. UDELSON: There were three readers for each patient who read independently. And then that was a rule set for creating the score. So two readers read the studies independently. If those readers agreed, in other words, if their reads were within two SDS units, a rounded average was used.

DR. NEATON: So the SDS was used for agreement as opposed to kind of the rest and stress separately?

DR. UDELSON: That is correct. The SDS was because that was the metric of interest. If the readers did not agree within 2 SDS units, a third reader was used. If two of those three agreed within 2, a rounded average of those two; otherwise, a small number went on to a consensus of the three readers.

So these were sort of a rule set prospectively put into place to combine -- "combine" is probably not the best word -- the three readers' scores into one score for the primary analysis.

DR. NEATON: I mean, I suppose in that situation there could be some information on the inter-reader variability by drug that could be important.

Did you look at that?

DR. UDELSON: Yes, we do.

DR. CARTER: We have those data, yes.

DR. NEATON: Let me just kind of -- while you're looking for that data, this may reflect my lack of understanding of this. But it goes back to what I think I heard, Dr. Udelson, you say, that in the literature there's a very strong -- the stress score is strongly prognostic with events. And so you've done a crossover study here. And in the two periods, you're doing a rest and a stress kind of test. And the rest tests are the same, essentially.

So one measure of whether things are kind of constant in this crossover study should be whether the rest scores are comparable during each of the two periods of your crossover study, which I presume you looked at and can kind of comment on as well.

But I don't understand why, if you're looking at equivalents, why you're compounding the error by subtracting off the rest score. Why not just look at the difference in the stress scores?

DR. UDELSON: Let me start with that, and then I'll go back to your agreement question. It's a very important point.

The summed stress score in prognostic studies is the most powerful predictor of outcomes because it combines --

DR. NEATON: The stress score?

DR. UDELSON: -- the summed stress score because it combines infarct and ischemia.

DR. NEATON: Right.

DR. UDELSON: Now, the summed difference score, the ischemia component, is really what is being generated by these drugs, on top of whatever infarction there is. We actually proposed at one point to FDA to use the summed stress score, and we were asked to use the summed difference score. However, you know, you're absolutely correct that it compounds the variability.

Can I have this slide up, please?

So these are the data. This was in the core presentation, number 78. So one point that you made was, one measure of the stability between the two exams is the resting score. So here, R is the analysis, the equivalent analysis of the resting scores.

Now, from a clinical perspective, if a patient has a small prior infarction at day 1, you know, day 7 they still have a small myocardial infarction. You know, the question is, how reliable or how variable is SPECT imaging of that infarct as read by independent readers with no clinical knowledge scoring segments -- you know, that segment might look like a 1 to me today, but a few days from now I might call it a 2 because it's sort of on the edge.

Probably the best measure is the computer analysis here because that gets rid of the variability of the human eyeball. So here are the rest scores and then the stress scores by themselves.

Can I have the next slide in the core?

DR. NEATON: Maybe while you're on that, I had a question about this slide, too, given this discussion.

I look at the confidence bounds here, and if my eyes are not playing tricks on me, it almost appears that the confidence intervals around the differences are smaller than the stress. And I'm puzzled by that.

DR. CARTER: Dr. Koch, yes, your perspective, please?

DR. KOCH: Gary Koch again. Well, your different score is basically subtracting a rest score from a stress score. And that can involve some reduction of variability because you're --

DR. NEATON: But you're taking a difference and a difference. And so I would have thought that would have --

DR. KOCH: Well, when you take the difference of the difference, then of course you're getting a contribution to variance from both the test agent and the referent agent.

DR. NEATON: So that's what the SDS is there? That's the --

DR. KOCH: Yes. But you have a -- when you're working with the rest score, you're working with a different of rest scores. When you're working with the stress score, you're working with a difference of stress scores. And when you're working with the difference score, you're working with the difference of differences.

DR. NEATON: No. That's what I thought, and I guess -- and again, my intuition would have said that those confidence bounds around stress score should have been narrower because all you're doing is subtracting off kind of a rest, which is on average the same for the two treatment groups in the two periods. And you're just adding unnecessary variation to the different statistic.

DR. KOCH: Yes. That indeed is the case. But you are moving towards more of a within-patient variance component. So you're getting two different contributions from a within-patient variance component through both the subtraction of the rest from the stress and then the subtraction of the adenosine from the binodenoson.

DR. NEATON: Do you have data that you can kind of show us after lunch on the -- just to quantify for us what the standard deviation of the stress scores were in the studies, as well as the standard deviation of the stress scores when adenosine was given twice?

Yeah. I think the sponsor -- so we can kind of basically --

DR. KOCH: Yes.

DR. NEATON: -- kind of use the same logic that you used in choosing the difference just to focus on the stress score?

DR. KOCH: Okay. Bring the slide up.

So this slide is essentially showing, for each of the studies, what the standard deviations are for the two methods, reader and computer, for each of the different arms that are being looked at. And this is pertaining to the summed difference score.

So here you see that the standard deviations are on the order of 7 in the first slide, which pertained to the reader. And basically, you're getting all of the arm, so you're getting the binodenoson and the adenosine, both when it was first and second.

DR. NEATON: Do you have the standard deviation of the difference?

DR. KOCH: Okay. So here we would want what would be the standard deviations that would apply to the summed difference score in a similar table.

DR. NEATON: I mean, I guess my question is --

DR. KOCH: This would be a table that would look just like the previous table. But maybe we can come back to that later.

DR. NEATON: I mean, I guess what I'd like to know -- because the literature that I saw, that you referenced, suggested that the stress score, as you indicated, was prognostically important and that, at least by my computations, you know, a difference along the lines that you kind of found here would be associated with roughly a 15 percent risk in cardiac events. And so we can argue about the clinical relevance of that. But I'd like to kind of nail that statistic and understand it for the committee.

DR. UDELSON: Can you put this slide up for a second?

This may be what you're asking for, the means and standard deviations of the summed difference score. And I think what you notice here compared to the previous slide is the standard deviations for the summed stress score is larger; there's a much wider range of values that happen when you have both ischemia plus infarction versus ischemia alone. So the standard deviations are wider for the --

DR. NEATON: No. I thought this was very helpful. And, actually, I looked at this, and I felt pretty reassured when I thought about the summed difference because your standard deviation differences for the adenosine kind of replication were very similar to when you gave binodenoson and then adenosine, and so that was reassuring.

I guess what I'd like to know is kind of what are the standard deviations or the differences in the summed stress scores and what the average differences on the summed stress scores are between the two treatment groups.

Because you've told us that -- and I kind of thought -- that's the way I interpreted the literature, that the stress scores are prognostically important. And in order to judge kind of the relevance or the differences, I'd like to kind of be able to kind of do the same thing you did with those stress scores.

DR. UDELSON: Okay. Slide up?

So here are the three studies, 305, binodenoson -- can you leave it up here, too, please?

The three studies. Now, this is the same analysis as we showed you of the difference in the summed difference scores, but now it's the difference in the summed stress scores.

Is this, Dr. Neaton, what --

DR. NEATON: I think so. I mean, just kind of of getting with you. So the average difference there --

DR. UDELSON: The top line.

DR. NEATON: -- is the top line, .66. And the 4.76 versus 5.32 is the standard deviations for the comparison.

DR. UDELSON: And the confidence intervals right below.

DR. NEATON: And 3.02 -- let me just kind of -- okay. So the standard deviations of the differences for the stress scores are all lower than the adenosine/adenosine comparison.

DR. UDELSON: That's right. And if we use the same equivalence boundaries that we had been talking about, plus or minus 1.5, all of these fall well within that.

DR. NEATON: Yeah. So that's the part where I guess this is help because do the same equivalence boundaries make sense?

DR. UDELSON: Well, I think we'd have to --

DR. NEATON: For the stress scores?

DR. UDELSON: I think we'd have to re-walk through our entire logic on page 51 there.

DR. NEATON: Right.

DR. UDELSON: No. Let me just make a general comment.

I think, we, as you might imagine, talked about this for many, many months in trying to come up with a clinical rational for something that's really very continuous. When you look at databases, mostly coming from the group at Cedars Sinai, of thousands of patients who've had SPECT imaging who are then followed for outcome events, there really is -- you know, there's sort of a discrete beginning when event rates start to increase, which is about 5 percent of the myocardium, from some of their studies. And after that, it's much more of a continuous scale.

So a delta, a difference of three, also seemed to make sense because it often will get you into different categories that are commonly used in the literature. But the rationale is contained in the briefing document there and we tried to think about this as rigorously as we could to make a cutoff for something that really is fairly continuous in populations.

DR. NEATON: And according to the paper that you referenced, a 1 percentage point difference in percent of myocardium, so the score divided by the total, is associated with about a four and a half percent increase risk of cardiovascular disease. And so I don't know how that corresponds to 1.5 standard deviation units in the stress score.

DR. UDELSON: I think at lunch I'll re-look at that part and try and get back to you.

DR. HARRINGTON: Jim, I'm going to move on because we've got an increasing list of people who want to jump in. I'm going to ask people to keep their questions brief, if possible. We will have time this afternoon to come back.

Lyle, we'll start with you, and then we'll go to Mike.

DR. BROMELING: I'd like to talk about the test accuracy of the two agents as measured by sensitivity and specificity.

You used a subset of those who tested positive and referred them to arteriography, right? Among those that tested positive and among those that tested negative, you used a subset of those.

It might bias these results for sensitivity and specificity in a general area called verification bias in statistics. However, there may be a way to improve that or get more reliable estimates of sensitivity and specificity if we knew the exact mechanism by which a patient is referred to arteriography.

Let me get this straight.

Did you just use the adenosine scores only to refer a patient to arteriography; or did you use other clinical and symptoms of the patient to refer them to arteriography?

DR. CARTER: Yes to the latter. But I'll ask Dr. Udelson to give you more precision.

DR. UDELSON: Thanks for that question.

As I pointed out, the referral to arteriography was not protocol-directed, number one. It was made by the patient's clinician on the basis of the clinical data that they had available plus the adenosine SPECT information that was read by the site.

So these patients were referred into a nuclear cardiology laboratory for an adenosine SPECT study by their clinicians. They were recruited into this protocol, had both studies in random order. The adenosine data were then read at the site by their local people, given to the clinician, the patient's clinician, who then made a decision.

So we actually do not know what the site reads were. We do not capture that data. And we don't know the weight, as it were, of the clinical versus the adenosine imaging data that led to the decision for coronary arteriography. But it was clinically driven, which then of course, as you totally correctly suggest, drives referral bias into those estimates.

DR. BROMELING: Yeah. There are statistical techniques for getting more reliable estimates of sensitivity and specificity. If you understand the mechanism, the referral mechanism, there are some statistical techniques. You can find those in Pepe -- there's a reference by Pepe's book and in Zhou's book, whole chapters about verification bias.

So that caught my eye. So you may be able to, you know, re-analyze that and get other estimates, more reliable estimates, of those two test accuracy measurements.

DR. HARRINGTON: Thank you.

Mike, I'm going to go to you, and then I'm going to stop the questioning so we can hear from the FDA. But we're going to have time to come back to it. So go ahead, Mike.

DR. DOMANSKI: Okay. I certainly appreciate the difficulty of designing studies here. But I'm sort of bothered, not by very technical statistical issues, but I'm sort of bothered by some overall issues that relate to how you design this study. And I'd like to lay them out for you, not so much as a challenge, but maybe you can kind of help me work with it.

The first thing is, you know, it's fine to talk about prognosis, and that's an important use of the nuclear study. But the other way these studies are used is to ask whether or not there is a suspicion of coronary disease, and that drives the decision for catheterization. And that's a common clinical use of this.

If you look at your Studies 301, 302, and 305, using adenosine as the gold standard that you put forward, then, in fact, if one just looks at how many times you miss, what percent you miss, -- I calculated 27 -- you call a study -- abnormal studies were called normal in 27 percent, 22 percent, and 24.5 percent of the time.

Now, that's mild/moderate/severe. And I

understand there are differences in prognosis and so forth. But if the absence of ischemia is what keeps a patient out of the cath lab, then it seems to me one is missing too much to use the test in that way, number one.

Then, secondly, when the concordance isn't what you want, we get a discussion of angiograms, and you take your gold standard and you go after your own gold standard with probably hopelessly biased angiographic selection. I don't think you're ever going to be able to sort out precisely why a clinician decided to send somebody to angiography.

So I'm bothered. I'm kind of bothered by the moving gold standard, and I would certainly stipulate to the fact that the results with the adenosine were unimpressive.

But that said, it just seems like the whole design shifts as the answer you want doesn't appear. So maybe you can help me with those things.

DR. CARTER: Let me just start by saying that the objective of the design was to look for agreement between the test, which was binodenoson, and the reference product, not a priori versus the gold standard, which is angiography. That information obviously allows us to calculate measures of accuracy.

So having selected adenosine as the reference agent -- because it's the most widely used myocardial perfusion stress agent in use in the U.S. today -- then it was incumbent upon us to show that the comparison and agreement, concordance, between the ability of the test and the reference to provide clinically important information was equivalent.

I sympathize with you and other members of the committee, and obviously with FDA, relative to the fact that we changed or we amended our analytical plan and our protocol. But we did so based on what we believe to be very good and sound reasons.

DR. DOMANSKI: Well, let me just -- and then I'll stop. But what bothers me is not so -- I'm not trying to get into a technical thing of, oh, you said this, and how you did that. I'm really looking at how many times you miss. And you miss a lot. You miss almost a quarter of the time. And that's the thing that I'm kind of trying to struggle with and make it look like concordance with the statistics. But I'm having trouble.

DR. HARRINGTON: So, Mike, that is going to be one of the key points for the questions that emerge this afternoon. So I'm going to allow the sponsor to stop here because we are going to come back to that very issue. That's one of the key things FDA wants us to discuss.

There's a list of questioners. We're going to keep the list. We'll come back to it right after lunch. But why don't we move on with the FDA presentation so we can hear another perspective on the data. And we'll start with Dr. Marzella.

DR. MARZELLA: Good morning. My name is Louis Marzella, and I will provide an overview of FDA's interim observations from the NDA review.

I will begin by restating the sponsor's major marketing proposal, and will then focus on the Phase 3 clinical study. So I will first summarize the study's development, discuss the key design aspects, introduce the efficacy data, and finally, I will summarize the safety data. Following my presentation, my colleague, Dr. Mark Levenson, will talk further about the efficacy data.

So as you've heard already, binodenoson is proposed for marketing as a dose regimen of 1.5 micrograms per kilogram injected intravenously over 30 seconds. As noted earlier, the proposed indication emphasizes the role of binodenoson as an adjunct to diagnostic myocardial perfusion imaging. Hence, the major efficacy outcomes in the Phase 3 studies were radionuclide perfusion images.

Again, as has been discussed earlier, this is an overview of the study designs. A major Phase 2 study, Study 206, established a paradigm that the sponsor carried over into the Phase 3 study development.

As shown here in this crossover paradigm, all patients had two sets of myocardial images. One set of resting stress images was obtained with binodenoson. The other set was obtained with adenosine. The sequence of stress agent administration was determined by randomization.

This slide sort of summarizes the key

features that were shared by the main Phase 3 studies. All three studies used multicenter, randomized, crossover designs. An important feature was that the image sets were obtained in a double-blinded manner. However, the study drug assignment was unblinded following completion of the last image set to permit patient management.

Given this unblinding, the adenosine image sets influenced the selection of patients who underwent further coronary diagnostic tests.

As noted in the second bullet, all images were assessed at a central reading facility, where readers were blinded to clinical information. And has been discussed thoroughly by the sponsor, a standard 17-segment cardiac perfusion model was used. And I will not dwell on those details.

Then following the assignment of scores to each segment, the scores were summarized for the rest and stress images to yield the SRS, or summed rest score, the SSS, or summed stress score, and the difference between these two scores, namely, the summed difference score or SDS.

Following the image evaluations, patients were followed for 60 days to collect information on adverse events and cardiac interventions, including coronary arteriography.

So for simplicity, we are referring to the Phase 3 studies by the acronyms Study 301, 302, and 305. The studies were performed sequentially, although the designs were finalized before the completion of Study 301.

In Studies 301 and 302, patients were randomly assigned to the binodenoson or adenosine administration sequence. An important difference in Study 305 was that in addition to these two sequence options, a third option was included, in which patients underwent two sequential adenosine administrations.

Now, you've heard already about the challenge posed by the original and modified primary endpoints. This is a restatement of the original primary endpoints.

Originally, all three studies had prespecified endpoints that assessed the correlation of each image set using a weighted kappa statistic. To address this endpoint for Studies 301 and 302, each binodenoson and adenosine image set was assigned to one of four possible SDS categories, which ranged from normal perfusion to marked perfusion abnormalities.

And the success criterion for the weighted kappa value, required at the lower limit of a two-sided, 95 percent confidence interval, exceed 0.61. This value was chosen by the sponsor based on the results of the major Phase 2 study results, and on certain results previously obtained with adenosine and available in the literature.

Now, as to the analytical modification, Study 301, the first of the three studies, failed to achieve it correlation endpoint, conceivably due to underestimated sources of variability, of which there are many, such as test/retest, physiologic variability, and image interpretation variability.

At this point, Studies 302 and 305 were ongoing, and it became apparent that the design of these studies were such that the specified primary endpoint correlation would likely also not be achieved.

Subsequently, the sponsor changed the analytical plans for Study 302 and 305, before unblinding the image data. The new primary endpoints were defined as noninferiority comparisons of the average binodenoson SDS to the average adenosine SDS, with the noninferiority margin described as containment of the 95 percent confidence interval for the adenosine minus binodenoson difference between minus .15 and plus .15 units.

Now, we at FDA had several concerns about this primary endpoint, as highlighted here. And notably, the modified primary endpoint compared average scores across each population of image sets, not pairwise comparisons of each image set. And the concern that we have is that this approach may lessen the ability to verify agreement of individual sets, since these sets do not undergo pairwise comparisons.

A population approach, as one may use for sensitivity or specificity estimation, may be useful if the populations are defined by an established truth standard. However, the use of a reference test without a truth standard may increase the variability within the study assumptions, thereby inappropriately increasing the likelihood of achieving the desired noninferiority of a new test agent.

In addition to the population approach, we had concerns about the robustness of the data used to develop the newly defined noninferiority margin.

To briefly consider additional endpoints, these are as listed here. There were multiple other summed perfusion score comparisons, and importantly, there was also a use of MPI scores in patients where coronary arteriography was used as a truth standard, and MPI scores where prespecified clinical outcomes were used as a truth standard. Additionally, as the sponsor discussed in detail, multiple safety endpoints were prospectively defined.

As to the major eligibility criteria, all studies enrolled adults able to undergo pharmacologic stress MPI, and the enrollment was targeted toward prespecified proportions of payments with coronary artery disease likelihood defined as low, immediate, or high. And since all patients were to receive adenosine, the eligible patients had to have no contraindication to adenosine, such as reactive airway disease.

Now, let's begin to consider the results. The patient disposition is summarized here. Overall, 1,354 patients were randomized. And as shown in the secondary row, in the second row, approximately 90 percent of the patients were included within the efficacy population. And we considered this to be a relatively successful proportion. However, coronary arteriography was performed uncommonly in the studies, with only 15 percent of the overall population undergoing the procedure.

Turning over to baseline characteristics, the patients' major baseline characteristics were relatively similar across the studies, with the median age approximately 63 years and with women accounting for slightly more than half of the population. The coronary artery disease likelihood varied modestly across the study, with the intermediate group appropriately predominant within each study.

This slide shows the primary endpoint

results. And these are summarized here in terms of the modified and original definitions.

Shown in the first row, success with the modified endpoint was shown in all studies based upon confinement of the confidence intervals within the desired limits. As shown in the bottom row, the weighted kappa values were below the desired 0.61 limit within all three studies. Therefore, success was not achieved on any of these correlation endpoints.

Again, as previously noted, coronary arteriography was the truth standard for the study drugs MPI performance characteristics. And shown here are the average sensitivity and specificity values within each of the studies. And in this table, binodenoson is identified as B and adenosine as A.

In general, the estimates were relatively variable across the studies, with adenosine tending to have higher sensitivity but lower specificity compared to binodenoson. However, as has already been highlighted in the discussion this morning, the meaningfulness of these data are questionable since only 16 percent of the patient population is included in the analysis, and the MPI results influence -- and other unknown factors influence the decision to perform coronary arteriography.

The study also prespecified certain follow-up outcomes as important clinical endpoints to potentially also serve as truth standard for comparison to the image sets. However, only 6 percent of the patients experienced these outcomes, and this number is, of course, too small to meaningfully estimate sensitivity and specificity.

The clinical endpoints consisted mainly of coronary revascularization procedures. No deaths occurred in the studies. Conceivably, the relatively low number of events was due to the limited follow-up duration of 60 days, as well as the potential impact of the MPIM blinding upon the performance of the coronary arteriography.

Now, turning over briefly to summarize the safety data, within the entire development program, 1,674 patients were exposed to binodenoson, and of these, 1,166 were included within the Phase 3 safety database. In describing the occurrence of the adverse events, the events were assigned to study drug periods based upon whether they began after binodenoson or adenosine.

Overall, adverse events were experienced by over 90 percent of the Phase 3 patient population, with the numeric rate higher for adenosine than binodenoson. Similar proportions of patients discontinued the study drugs because of adverse events, and similar proportions also experienced serious adverse events.

The grading of adverse events showed that most events were mild to moderate in severity, and the numeric proportion of moderate to severe events was lower in the binodenoson period compared to the adenosine period.

A series of adverse events were prespecified as ones of special interest in the studies because these events are related to pharmacologic effects of the study drugs. And this table shows the pool results from these analyses.

Overall, the numeric rates of 7 of these 11 events were lower in the binodenoson period than in the adenosine period. The 4 other adverse events, highlighted here in this slide, largely occurred in similar proportions between the two study drug periods.

Of particular note is the bottom row, where second- or third-degree heart block was reported to have occurred among no patients in the binodenoson period, but 27 patients in the adenosine period. And these data are consistent with the sponsor's postulate that specificity of the product for specific adenosine receptor subtypes is different than that of the reference product.

The times from start of study drug administration to onset of the most common adverse events for the majority of patients in each treatment group were within zero and 10 minutes, again, consistent with the pharmacokinetic profile of the drug. A few adverse events were observed to begin beyond one hour after administration of the drug, and the duration of these events was generally brief, with medium duration times less than 10 minutes.

This slide summarizes another adverse event of clinical importance, which is due to the activation by the drugs of adenosine receptor present systematically other than in the coronary artery circulation.

So with regards to hypotensive events and heart rate changes in the pool study, overall, the conclusion is that these changes occurred at similar proportions between the two study drug periods.

Now, the clinical protocols included relatively detailed plans for multiple other safety outcomes that largely assessed symptom tolerance. These outcomes included visual analog scores of a specific adverse event's intensity, estimates of symptom bother scores, as well as a summary of patient preferences regarding the study drugs. Overall, the pattern of these outcomes favored binodenoson.

So then let me summarize. Our preliminary review indicates that the safety data revealed no unique concerns for use of binodenoson as a pharmacologic stress agent. However, the efficacy data are much more challenging, particularly because the studies were not designed to sufficiently assess test/retest variability, which is one of the known challenges with image-based clinical studies.

Conceivably, this variability may have contributed to failure of the studies to achieve the original image-set correlation primary endpoints. In anticipation of this unsuccessful correlation, the primary endpoint for two of the three studies was modified and average summed difference scores were compared. This modified primary endpoint was achieved in all three studies.

At this point, I'll call on my colleague, Dr. Mark Levenson, our lead statistician, to discuss further the efficacy analysis and data.

Dr. Levenson?

DR. LEVENSON: Good morning. My name is Mark Levenson. I'm the primary statistical reviewer for CorVue. Today I will address the confirmatory studies for CorVue and the level of evidence they provide for efficacy. I'll try to briefly go through the material that you've already heard today. First I will review the design endpoints analyses for the studies. Then I will present some key efficacy results. Finally, I will discuss the results.

There were three confirmatory studies for CorVue, 301, 302, and 305. Studies 301 and 302 were crossover designs. Each subject underwent one adenosine session and one binodenoson session. The order of the two sessions were randomized. Half the subjects received adenosine in the first session and half the subjects received binodenoson in the first session.

Study 305 differed from 301 and 302 in that one group of subjects received two sessions of adenosine. This was the only study that had a within-study measure of adenosine variation.

As you've heard, for the image evaluation, each rest/stress image pair was reviewed by two central blinded readers. From their review, the summed difference score or SDS was calculated. SDS can range from 0 to 68.

The agreement between adenosine and binodenoson was evaluated based on SDS. The original agreement measure used the kappa concordance statistic. The kappa statistic can vary from minus 1 to 1, in which 1 represents perfect agreement and zero represents agreement equivalent to chance.

The kappa statistic was based on four categories of SDS: 0 to 1, 2 to 4, 5 to 8, and greater than 8. The kappa statistic can be understood with a cross-tabulation. For each subject, the SDS categories for the two drugs are cross-tabulated, as seen in this table.

Kappa is high when subjects have the same SDS categories for both drugs. That is, subjects generally fall on the main diagonal of the table. Note that the kappa statistic is dependent on the prevalence of subjects in the categories.

The revised agreement measure used the mean difference in SDS between the two drugs. Here is an example with dummy data explaining this. The first subject had an SDS of 3 for binodenoson and 5 for adenosine. The difference is minus 2. The second subject had an SDS of 5 for binodenoson and 3 for adenosine. The difference is minus 2. The third subject had the same value of 2 for both drugs, and the difference is 0. The fourth subject had a difference of 0.

The mean difference is .25. You can see that the mean difference in SDS is subject to cancellation where the mean can be small even if the individual subject differences are not small. The mean of the difference in SDS is equivalent to the difference of means of SDS. Therefore, the measure is more of a population summary than a summary of individual subjects.

The original success criteria for 301 and 302 were based on kappa concordance or four categories of SDS. For Study 305, the original success criteria was based on kappa concordance of four categories of overall clinical interpretation. For all studies, the success was defined as kappa exceeding .61 or, more precisely, the lower limit of the 95 percent confidence interval of kappa exceeding .61.

The revised success criteria for all three studies had two components, the mean difference in SDS had to be less than 1.5 units; in particular, the 95 percent confidence interval had to be within plus or minus 1.5.

The second condition protected against extreme cancellation. The second condition was that less than 10 percent of the patients had extreme discordance between the two drugs; for example, one drug having an SDS of 0 or 1, and the other drug having an SDS greater than 8. The second criteria does not protect against other disagreements in SDS categories.

As we have heard, cardiac angiography was not required in the studies. Information was obtained for subjects that underwent the procedure within 60 days. It is important to note that the images were locally unblinded after the second image session for subject management. This likely affected the angiography sample.

Now I will present the key efficacy results from the three studies. First, the original success criteria.

Here is the SDS concordance table between binodenoson and adenosine for Study 305. 197 of the 391 subjects in the efficacy set had an SDS of 01 for both drugs. This represents 50 percent of the subjects.

Looking along the main diagonal, 236 subjects, or 60 percent, had agreement in the two drugs in these SDS categories. The majority of these subjects were in the 01 category.

Twelve subjects, or 3 percent, had extreme discordance, that is, 01 for one drug and greater than 8 for the other drug. Fifty subjects, or 13 percent, differed by at least two categories in SDS.

Here are the kappa estimates for the three studies. For Study 305, the kappa for the binodenoson agreement and the kappa for adenosine/adenosine agreement are given. Recall that the success criteria was that kappa exceeded .61. In no study did the point estimate for kappa exceed this value.

For Study 301, the value was .25, for Study 302, the value was .36, and for Study 305, .43. The lower limits of the confidence intervals, which were the basis of the statistical tests, were naturally even lower. Note that the kappa agreement of adenosine with itself did not achieve this threshold.

Now I will discuss the revised success criteria.

Here is a histogram of the difference in SDS for the 391 subjects in the efficacy set for Study 305. The differences range from minus 16 to 22. Twenty-six percent of the subjects, or more than one quarter, had differences beyond plus or minus 3, represented by the dashed lines. Five percent of the subjects had differences beyond plus or minus 9.

For all three studies, the revised success criteria were achieved. The confidence interval for the mean difference for the three studies fell within plus or minus 1.5. The confidence intervals for the percent of subjects with extreme discordance were all less than 10 percent.

Now I'll briefly discuss the angiography results. The sensitivity and specificity of binodenoson was near 67 percent. The sensitivity of adenosine was higher, and the specificity was notably lower. This may be due to the select sample who underwent angiography. Likely, the decision to proceed to angiography was based on the results of the approved agent, adenosine.

Here we can see among the subjects for cardiac angiography procedure there was a higher percentage of positive results for adenosine than binodenoson, 63 versus 53 percent. Thus, the sample is over-represented by positive adenosine results.

Now I will discuss the results of the studies in the context of providing statistical demonstration of efficacy.

First, the limitations of the designs. Studies 301 and 302 did not contain a group of subjects that received adenosine in two sessions. Therefore, the adenosine/adenosine concordance could not be measured. In fact, you can get perfect adenosine/binodenoson concordance by finding normal perfusion in every image.

We saw in study 305 50 percent of the subjects had an SDS of 0 or 1 for both drugs. A noninferiority design with an adenosine/adenosine arm would enable some assay sensitivity. This type of design was used for the confirmatory studies for regadenoson.

Here I present the noninferiority analysis of Study 305, the only study with an adenosine/adenosine arm. In the noninferiority analysis, the binodenoson/adenosine kappa is compared to the adenosine/adenosine kappa.

Looking at the point estimates, the difference in kappa is about .1, .43 versus .53. The confidence interval for the difference goes down to minus .24. The value of .24 is likely too large for a noninferiority margin. A larger sample size may have reduced the width of the confidence interval to fall within a reasonable noninferiority margin.

Now I will discuss the limitations of the analyses. As we have seen, a small mean difference in SDS across patients does not imply small differences in SDS on the patient level. In fact, it is possible for every patient to have a different diagnosis for the two drugs and still have a mean SDS difference of zero.

Twenty-six percent of the patients in Study 305 had an SDS difference greater than 3. The mean SDS difference is not an acceptable endpoint from a statistical perspective. There were substantial differences in results based on kappa and the SDS difference.

My final comments, the original success criteria failed for all three confirmatory studies. The revised success criteria are inadequate. The angiography results are based on a limited subject sample and are potentially biased. In any drug approval, the demonstration of efficacy is based on prespecified and adequate primary analyses.

Finally, I conclude that efficacy has not been statistically demonstrated for CorVue. Thank you.

DR. HARRINGTON: Thank you.

Dr. Levenson, maybe we can have you stay up there and we can start with the statistical questions to the FDA, and then we'll have you come back as needed.

I'm going to start off here. I was most interested -- well, there's a lot of things I'm interested in here, including why the sponsor hadn't followed your advice on a design, but we can come back to that maybe this afternoon.

On slide 17, when you note the one trial when there's actually an adenosine/adenosine comparison, and you note the kappa was .53 with the associated confidence interval, in your view what does that tell us about adenosine?

I mean, I've jotted down at least four things here, one of which is that adenosine behaved poorly in the experiment. The other is that there's just not been sufficient testing of adenosine; it was only the one study, as you've pointed out.

The third is that adenosine is problematic in and of itself, in which case it raises a lot of other questions about the appropriateness of the comparison.

Then the final is that the test statistic, the kappa test is not appropriate for what we're trying to accomplish here.

So could you comment?

DR. LEVENSON: Sure. I'll start with maybe a more straightforward answer.

Adenosine-adenosine -- adenosine is

obviously a variable product, as we've seen throughout the day. When you repeat the procedure, you may get different results. And as we see in this slide, adenosine couldn't meet the concordance that the sponsor was trying to achieve for their agreement with it. So in effect, it was an impossibility, as the sponsor has pointed out.

If agreeing with a drug that doesn't agree with itself, I'm not sure how valid a success criteria that would have been itself. So using adenosine as standard of truth can be problematic.

DR. HARRINGTON: I don't know if Dr. Marzella or Dr. Rieves want to help us out here. But this is going to be in part the essence of the day. Right? That if, to use the vernacular, adenosine's a bad comparator, how do we judge relative to that?

DR. LEVENSON: Well, larger sample sizes would have --

DR. HARRINGTON: Dr. Marzella or Dr. Rieves, do you have a comment on my question?

DR. RIEVES: Well, the first thing as to whether or not it's a bad comparator, it's an acceptable comparator because it's on the market. You know, on the face of it, it's an acceptable comparator in the consideration.

So I think -- you know, I catch myself. I think there are a number of lessons to be learned from the regadenoson experience, for example, where the primary endpoint -- the concordance was achieved on that primary endpoint for that product.

But there was also success demonstrated in multiple other concordance assessments. And Dr. Tony Mucci reviewed that, has a nice review of that, may comment there. But here, with regadenoson, he showed that that agreement was very consistent on multiple endpoints throughout.

So yes, we do. It's an approved product. Its out there. It's obviously clinically used. It is a challenge, though, as we see today.

DR. HARRINGTON: So let's go to John, Sanjay, and Jim.

DR. FLACK: First, I need to try to understand the validity of trying to, you know, break this continuous measure up into all these quadrants and then show agreement. And we've talked a lot and we've quizzed the sponsor about the validity of doing certain things. And I do have problems with the SDS score.

But when we go back and look at these categories and say -- I mean, I need to understand what the validity of that is. It tests either normal or abnormal. And I didn't see anything that just basically showed agreement between normal and abnormal.

So these finer gradations I'm troubled with. And also, I'm bothered by the fact that there's a problem to me with just using gold standard of 50 percent coronary blockage. I mean, it's implying that that is the route to coronary ischemia.

And you have to have an anatomically visible lesion to have coronary ischemia, and you've got a functional test that might be picking up ischemia that's not mediated through an angiographic block. So the block may be associated with ischemia or not associated with ischemia.

So I wondered, were there any sensitivity analyses using various cut points? But I'm really bothered because a ventricle can be ischemic and the nuclear test can be right, and we're using something that seems so retro, just anatomic obstruction of a vessel at 50 percent.

As a non-cardiologist who knows a little bit about the coronary circulation, that really bothers me. And so when these tests don't agree with the angiogram, I'm not so sure that I know which one is necessarily right because even in the myocardial defects in the nuclear scans are not even matched up to the region of where the block is.

It's just sort of a coronary population. You've got a block. You have ischemia. It doesn't have to be in the same region, and so I'm really confused here.

DR. HARRINGTON: So I think, John, I think this is also going to be something we come back this afternoon. But I think it also gets to Jim Neaton's question a little while ago, which is that the summed stress score, as to what it offers versus the difference score.

I thought Dr. Udelson did a reasonable job of trying to explain that to us, that the summed stress score is in fact a powerful prognostic indicator. And you're absolutely right, John. That doesn't mean you have fixed obstructive coronary disease if you have an abnormal summed stress score.

So your point is well taken. And whether or not the truth standard should be coronary arteriography is probably a longer discussion.

Let's to go Sanjay and then to Jim.

DR. KAUL: I had one comment about slide 27. You said that noninferiority design would enable some assay sensitivity. And I'm not quite sure because if, as we just discussed, adenosine is not a valid reference control, all you're going to demonstrate is that it's as effective or as ineffective as the adenosine.

I have one question on slide 17. If we do accept adenosine to be a valid internal control, there was only one study that had a valid internal control, and, ideally, would have liked to see that in Study 301 and Study 302.

But if you take the adenosine/adenosine kappa estimate as the comparator and do a comparison against that with Study 301 and 302, you can see that there's no overlap whatsoever with 301, and there is minimum overlap. And if you do a statistical comparison, would it be fair to say that it is statistically inferior, the binodenoson/adenosine estimates in 301 and 302?

DR. LEVENSON: Okay. I don't want to -- I can't comment on whether it would be statistically inferior. I would be concerned that the adenosine/adenosine variation is coming from a different study.

But the slide I left up there does do the noninferiority for 305. So you do get confidence intervals. The very bottom of that slide, the .4 to .65, does represent the confidence interval on -- wait --

DR. HARRINGTON: You probably are not looking at slide 28.

DR. LEVENSON: Yeah.

Can we go to slide 28? Okay.

So the bottom here, this confidence interval from minus .24 to .05 does represent our statistical estimate of the difference in kappas within Study 305.

DR. KAUL: So if you were to translate that into percent agreement, as was done with the regadenoson program, what will this translate to? I'm trying to see if we can compare the two.

DR. LEVENSON: Well, regadenoson used a very similar design to what this is trying to do here. And their noninferiority margin -- well, they did not use kappa. They used a statistic similar to kappa. And they provided some justification of what the equivalent kappa noninferiority margin would be, and that was .2.

So that .2 -- the fact that this goes down to minus .24 means they would not have met the noninferiorities setup in regadenoson. But there are other differences as well. You know, regadenoson was not directly based on SDS. It was based on the number of reversible segments.

DR. KAUL: Thank you.

DR. HARRINGTON: Thanks, Sanjay.

Next is Dr. Tatum.

DR. TATUM: I wanted to go back to this issue because I know there are other imaging agents that are on the market that don't perform very well. And in other discussions related to those, I had understood that in that case, an equivalency would not be sufficient for approval.

Have we changed our position on that or is that still the case? I have the agent in mind, but I don't want to say what it is.

DR. LEVENSON: I wish I knew which agent. I can't deliver on the specifics.

DR. TATUM: It's an oncology agent.

DR. LEVENSON: I don't think we have changed. The use of a reference agent is -- as articulated in the 2004 guidance, we regard it as reasonable. It can, of course, be a challenge to demonstrate, as we're seeing here today.

But I think it would be difficult to discount the use of a product that is on the market and that is widely used, in fact, as a reasonable comparator. Of course, it leads to all sorts of analytical and logistical challenges. But I don't think we would be viable in discounting a problem among the most commonly used agent.

DR. TATUM: But let's, to use a term, ratchet the dialogue a little bit differently. And that is, in the case where you're dealing with a product where there's potentially some problems or it could be superior and you're looking at a comparator, you would expect it to perform at least as well or better and not to be -- your downward piece is very, very low at that point.

How much worse could it be? What is the significant benefit? We're really challenged. Where's the significant benefit?

DR. LEVENSON: Your point's exactly right. And that choice of the goal and the claim, if you will, the diagnostic claim, we leave that largely in the sponsor's court, you know, based on their molecule, what they expect.

But you pointed out one of the limitations in using a reference agent, what if the new agent's better? We have a challenge.

DR. HARRINGTON: Go ahead, Jim.

DR. TATUM: One other comment. I wanted to go back to the angiography and the ischemia and everything. All of us know that when you're using vasodilators, you rarely induce ischemia. We're talking about changes in vascular reactivity here. And the anatomical correlation may or may not work. I've seen plenty of cases with horrendous perfusion scans which they said, gee, there's no significant disease, until you went in with a Doppler wire.

So, I mean, that's really where we're talking about the gold standard. And I do have some concerns, which we can talk about with the sponsor later regarding that.

DR. HARRINGTON: Yeah. I think we're going to have, given a short public hearing session, some time for additional questions to the sponsor.

Let's move to Dr. Black.

DR. BLACK: I want to talk about -- some of my confusion, I think, is reasonably similar to John's. I'm not a cardiologist, but a little familiar with the disease and how we interpret it.

I feel almost like a patient representative here in that I may be getting one of these things soon and need one, and I'm not sure I'd want to get any of the things that you've shown me, including angiography.

So I would need some help. And I think I'd rather talk now than a little bit later as to what the agency would expect. You did approve something recently, which I wasn't familiar with. And I think we have a drug which seems to have a safety advantage. I don't think anyone's argued about that. And if so, I'd like to hear what the argument was.

It certainly seems to be better tolerated, and I think they certainly did a decent job with that. And I think there doesn't seem to be any agency disagreement. But the fact that it's on the market and it's been on the market for a decade or more, but it seems to have some problems, how do we deal with those?

DR. LEVENSON: I wish I had a simple answer. But as has been pointed out, the detection of stenoses leads to a claim usually for an imaging product along the lines of detection of obstructive coronary vascular disease. It does not usually lead to a claim of ischemia. Those are different topics.

Here we're particularly challenged in developing the pharmacologic stress agents because we are interested in detecting myocardial perfusion abnormalities, which, as we all know, is very different from obstructive pathology, that sort of thing.

So in a certain sense, you can see why the sponsor would choose a reference agent because what are the alternatives for myocardial perfusion? PET? We really don't have many options.

So I think we're in a dilemma. If anyone has any suggestions as to alternatives, we're delighted to hear it. But candidly, I'm not aware of alternatives.

DR. HARRINGTON: Sebastian?

DR. SCHNEEWEISS: I want to go back to the kappa statistic. And the disadvantage of the kappa statistics, which is just one number, is that it doesn't tell you where the disagreement comes from.

Is it random disagreement or is this systemic disagreement?

When we look at the cross-tabulation of Study 305, we see there's a systemic or there's a tendency towards a systemic disagreement in a way that bino is actually scoring lower than adenosine, which is then reflected in the sponsor's analysis of the SDS differences, where we see a negative .68, which actually reaches formal statistical significance.

I'm not talking about clinical significance here. It's really statistical, which makes Study 305 really interesting, I think, to explore what is different in 305 from 301 and 302, where we don't see that, where the noise seems to be random, as well as in the adenosine/adenosine comparison, where the disagreement seems to be randomly distributed across this cross-tabulation.

So my question is, to FDA and to sponsor, I guess, what is different in the study population of 305? From what I can see only is there's known CAD. And I really want to learn something from this. Can we learn something with regard to the test performance here by looking at the differences in the population? And known CAD can mean a lot of different things.

DR. HARRINGTON: So very good question.

Dr. Levenson or Dr. Marzella, do you want to comment as to what might be different about Study 3 other than the fact that we have the adenosine/adenosine comparison?

DR. MARZELLA: I think that the proportion of patients with the likelihood of coronary disease by design was slightly different. Other than that, I'm not aware of any other differences. Maybe the sponsor has additional comments?

DR. HARRINGTON: Dr. Udelson or Dr. Carter, do you want to weigh in on this? Then we'll go to Dr. McGuire.

DR. UDELSON: The targeted population was slightly different in the two trials, not the low or intermediate likelihood but the high likelihood --

Can I see that slide? Not this one, but the demographics of 302 and 305.

Well, just to jump onto that, in Study -- here we go. Thank you.

In Study 302 on your left, our targets and the actuals had 25 percent high likelihood and 25 percent know CAD. And Study 305 was slightly different in that there were 10 percent high likelihood and 40 percent known CAD as the targets and the actual.

Now, you may ask, why did that change? Before Study 305 started, in parallel to these -- can I see that next one that you were putting up -- in parallel to these trials, we actually performed a 5,000-patient observational outcome study of pharmacologic stress practice around the world.

This was 90 centers in five different countries who recorded data on pharmacologic stress patients, consecutive patients, in their lab who consented to enroll, and about 80 percent consented, over a 20-day period of time. So there were 5,000 total patients. And we did this to get sort of a sense of practice and patient patterns and referral to angiography around the world.

So here are the pretest likelihood categories in this outcomes trial, as it were.

So we had this while designing Study 305. And so we adjusted these high likelihood and known CAD to reflect these because we really went into this wanting to make the populations in these studies generalizable to the populations undergoing pharmacologic stress testing around the country and, indeed, around the world.

Next slide, please. Thank you.

Let me, if I could, Dr. Harrington, just go on a little bit because the targeted proportions completely drives the amount of ischemia. And, in fact, the amount of ischemia or reversible defects, as Dr. Tatum mentioned, was completely predictable by the population that we enrolled. And here are the distribution of ischemic abnormalities by these pretest likelihood groups in these 5,000 patients studied around the world.

Now, I might mention that part of this real life study was that we had the sites score -- the SPECT scans. This is not core lab analysis; this sort of real world amount of ischemia.

And what you can see, this is severe

ischemia, summed difference score greater than 8, moderate and mild, although I note we don't have non-ischemic in here. So non-ischemic is the rest of the bar.

So what you can see, for instance, in a patient with an intermediate pretest likelihood by symptoms, age, gender, by ACCHA criteria, only 30 percent have any reversible defects in the distribution, mild, moderate, severe seen here. And interestingly, even among patients with a high pretest likelihood, so a 60-year-old, typical angina, male, you know, about 40 percent have ischemia and only a small percent have severe.

So the distribution among the categories -- one of you made the point of the distribution -- this is what you get when you set up a study to reflect the population being referred for stress testing. If you want an angiographic population, of course, that's an entire different study, and then that relates back to Dr. Domanski's comments.

So there were some differences between 302 and 305. The overall prevalence and the distribution of the ischemic categories was actually not very different. So then, getting back to the actual question, I don't think that explains the slight difference.

DR. HARRINGTON: Thanks, Dr. Udelson.

Let me go to Dr. McGuire and then to Peter. We got a list of people who've been waiting.

DR. McGUIRE: Yes. A fairly brief question for Dr. Levenson about the kappa statistic that I'm not quite as familiar with as other statistical measures.

What is the influence of partitioning over 4x4 tables versus 2x2?

The reason I ask is while the graded associations with severity of SDS are certainly correlated with clinical outcomes, probably slightly less correlated with obstructive disease, and it's uncertain the relevance here when we're trying to compare two agents to demonstrate perfusion deficits.

In clinical practice we typically dichotomously categorize results from stress testing. There's normal or abnormal, and if abnormal, independent of the severity, tends to drive clinical decision-making.

So if you collapse these cells study to study, you get between 26 to 34 percent discordance with the results, which I find a little bit unsettling. And what influence collapsing these into 2x2 tables would it have on the kappa statistic using the same data?

DR. MARZELLA: I performed an analysis based on a binary categorization of kappa. And the categories I did in that were 01 versus all other categories. I don't know if that's clinically what you're interested in.

DR. McGUIRE: That's what I did as well. I think that's clinically relevant.

DR. MARZELLA: And the kappa statistic is completely depending on the categorization. But in this case, the kappa was very similar. So the number -- the actual estimates we got for kappa were very similar when you used up this binary kappa versus this 4x4 table.

DR. McGUIRE: Okay.

One quick follow-up question, maybe, for Dr. Rieves, is you set the stage today by talking about high agreement, and referenced comparisons, and I believe the terms were exactly the same results between the two comparators. And when you collapse these into 2x2 tables with 26 and 34 percent discordance, I'm just curious.

Is there a numeric value you give to the level of agreement that is considered acceptable?

DR. RIEVES: I'm not aware of a numeric for any one single endpoint. For example, with the product we approved last year, there was success on the agreement demonstrated across multiple endpoints; for example, wall motion, SSS, the designated primary endpoint. So it's the totality of the endpoints, if you will. I'm not aware of any single number, if you will, that defines success overall.

DR. HARRINGTON: So maybe I could take the prerogative here and ask the sponsor -- we've had a series of these questions now that John Flack has brought up, now Darren's raising, Dr. Levenson commented on, of thinking about the data in this binary way of normal/abnormal and have some discussion. I think Mike Domanski first brought it up this morning.

The other piece of that is something that Dwaine Rieves just mentioned, which is that there are other measures other than just perfusion that you might get at the time of pharmacologic stress, things like LV dilatation, lung uptake, et cetera.

And do we have any of that data from the two modalities? And maybe, if Dr. Udelson knows, we could hold that till after lunch and you guys could check on that for us.

So let me drop down to Peter.

DR. CONTI: One thing that's not been discussed very much is some of the technical issues associated with acquiring these studies, and the variations and the scanners that were used, whether patients were rescanned on the same scanner. Issues like attenuation correction could be a factor here in terms of identifying abnormalities. And so, some subanalysis of the segments that are more frequently, perhaps, identified as minimally abnormal could have some influence on the information that's being obtained.

The reason I'm bringing this up is because I was curious on your slide 17, the way you had that .53 as the adenosine/adenosine. And I don't know if the sponsor or FDA has this information, but compared to the literature, are there any retrospective analyses that have looked at this to come up with a kappa in the range that would have been desirable for the sponsor, in .6, in that ballpark, that would have qualified, let's say, that particular study as being acceptable?

To the extent that on the sponsor's slide 79 there's a fairly significant drift as well as variability of SDS to the left, if you will, on the chart, does that imply, perhaps, that there is a fair amount of not only inter-reader variability but potentially technical variability on the acquisitions of the study, and could this be resolved with larger patients, larger patient trials?

DR. LEVENSON: The .53 kappa and adenosine/adenosine in Study 305 was actually very similar to the kappa we saw in a similar arm for regadenoson for the last approval. So I think the sponsor made that point as well, that these kappas are in line with what we saw in regadenoson.

DR. CONTI: But is there any specific adenosine/adenosine comparisons in the literature beyond the regadenoson study? Have people looked at that and are there kappa statistics on that for comparison?

DR. LEVENSON: I'm personally not aware of any.

DR. HARRINGTON: Yes, go ahead, Dr. Carter or Dr. Udelson. This would be helpful because it gets to the question about how good do you have to be relative to the reference standard.

DR. UDELSON: Let me actually address your question about the technical variability because that's a very important point that could have played a big influence.

A tremendous amount of effort went into trying to maintain high-quality acquisitions that were similar between the two imaging sessions. A lot of time was spent in investigator meetings and with investigators. And it was specified that the sites were to use the same camera, isotopes, protocols, imaging times, and all of these times are recorded and reviewed, many by me, to make sure they matched so that there was no influence of time after injection to imaging.

No attenuation correction was used by any of the sites because at the time these studies were done, only a small percent of labs in the country were using that, so we thought that that would not be appropriate.

So anything that could be controlled was controlled. And I think in data that -- we do have -- just again, to the technical point, if I could, the readers rated, while reading the images blindly, the quality of the images, and that should come up in a minute.

But in a nutshell, just in terms of interpretable or problematic for an uninterpretable, and you can see that, well over sort of 90 percent of both binodenoson and adenosine images. They're similar within the readers in terms of the interpretability of the images.

DR. CONTI: If I may, I mean, I guess the issue is not so much where it's interpretable or uninterpretable. It's a matter of whether the readers are calling abnormalities when there aren't any. And a situation without attenuation correction becomes very difficult. And so looking at your subsegmental analysis, for example, in the inferior wall or the anteroseptal regions of the heart, whether there are very commonly attenuation corrections issues and artifacts, might have some influence on your overall statistics.

One of the things that was not shown here was weight as a factor and certainly the weight of the patient. It does say gender but it doesn't say how large the breasts are. So these types of issues could have a potential influence on the interpretation of the data.

DR. UDELSON: They certainly could. And, of course, since each patient had both studies, theoretically it should influence it similarly in both cases. But we thought about these issues. And again, attenuation correction was not done at many of the sites

DR. HARRINGTON: Thanks, Dr. Udelson. We've got Dr. Paganini, and then we'll go over to Dr. Neaton.

DR. PAGANINI: My question is really for FDA. I guess our role here is to look at safety and efficacy. Safety seems to be reasonable. It certainly doesn't show any signal that this is adding unsafe to the patient population at all. But then it leaves us with efficacy, and then I guess I go back to the SPECT MPI.

Was that approved as a diagnostic or as a prognostic? And the rationale behind that is it was used as if you have a certain number or certain change, then it's prognostic, especially above a certain thing of a higher or worse outcome.

So are we using a surrogate as our standard? And then we're using a surrogate comparison as well. So are we in a situation where we're using a surrogate to compare a surrogate to something that's a surrogate?

DR. RIEVES: Dr. Paganini is hitting on something that we see fairly consistently in imaging product development, and that is that claims are frequently not geared toward a diagnostic use. They're generally a bit more subjective.

For example, the approved products are approved for use in myocardial perfusion imaging. They're not specifically approved for prognostic use, risk stratification, if you will. That degree of specificity is not within their approved indication.

As is true for so many of our imaging products, they are approved as tools. And how they're actually implemented in clinical practice is discretionary, is at the judgment of the using physician.

DR. PAGANINI: So if I could follow up one little piece, then. So therefore, the role this committee would look at the efficacy of this drug compared to the comparator drug in the confines of a surrogate-type study?

DR. RIEVES: Yes, sir. That is correct.

All imaging outcomes are surrogates. You're exactly right.

DR. HARRINGTON: I think, Emil, that Dr. Udelson showed us some interesting data this morning, one of which is something that Jim Neaton had already pointed out, that the SSS is related to prognosis as the SDS, which is related to referral to the cath lab. So clinicians are using these two markers, if you will, in different ways.

DR. PAGANINI: I would say, yes, that was very enlightening. And being a renal guy that hangs in the bathroom and not necessarily in the cath lab, one of the things that sort of intrigued me was this prognostic/diagnostic study that had such a poor predictability, regardless of what you use.   And I'm wondering if the patient population itself, they eliminated people with significant left ventricular dysfunction and also class 4, New York heart classifications.

Is that a standard elimination from this type of test? I would think those are the people that can't run three miles.

DR. HARRINGTON: Well, I think that we can get into this after lunch. But I thought the registry data that Dr. Udelson showed at least suggests the way that the test might be being used in clinical practice, and also points out the fact that people do with the test very different things. There's not a straight line from test to the cath lab that necessarily is a logical one.

DR. PAGANINI: And one little quick thing. Was there a 304 and a 303? Is there just a 301, 302, 305? What happened to 303 and 304?

DR. HARRINGTON: Man, you bathroom guys are pretty smart.

[Laughter.]

DR. PAGANINI: I know, you know, if they start and then they stop and then they start again, I know they have a problem.

DR. CARTER: It sounds a bit like the development plan and the timeline that I showed you. There was just a 301, a 302, and a 305. And I wouldn't worry too much about the numbering.

DR. HARRINGTON: It must almost be time for lunch. But go ahead, Dr. Neaton, back to Frank, and then Dr. Krantz.

DR. NEATON: So I thought Mark Levenson gave a very nice talk in terms of kind of highlighting a couple things. I mean, there is the problem with looking at the average difference. There's problems with the kappa, too, which I think have been highlighted, and I think I'll point out a major omission from the designs, which is getting more data on the concordance of adenosine and adenosine in the trials.

However, in fact, while I think it's a limitation, an important one, I'm not sure exactly yet in my own mind how I would use the data because -- and so I guess one question I have for the sponsor is, where did the .61 come from? I mean, I saw no rationale except a reference, I think, to you, Dr. Koch. And so I just didn't understand where the number even came from in terms of clinical relevance and importance of agreement.

DR. KOCH: Gary Koch again. My impression is that the .61 was partly motivated by the 206 study because, as you recall, the 206 study had a fairly high kappa, and the lower confidence limit for that study was high as well.

You can go ahead and put the slide up.

DR. NEATON: I mean, that just kind of says to me that, well, okay, you chose a number you thought you would hit. But what's the relevance of it?

DR. KOCH: Well, I mean, first of all, with kappa, it's very challenging to get high values of kappa. You have to have very tight agreement in the categorical variables. So that's why a paper I did some time ago noted that you would have fairly substantial agreement if it was above .6 just because the within-patient variability has to be very small. You have to have virtually everybody on the main diagonal and a few people on the two diagonals that are to the right and to the left. And you saw that in 206.

DR. NEATON: Yeah. I mean, I agree with that. And I think my comment would be, looking at the 206 data, I think kappa seems like going down the wrong path. But somehow, when they chose to do the -- but the .61, then, had nothing to do with assay sensitivity or kind of any earlier data relating to either predicting going to the cath lab or predicting clinical events.

DR. KOCH: Well, again, if you have .61 as the criterion for a lower confidence limit, recognizing the variability in estimates of kappa that come from a sample, you would potentially need to get an observed estimate of kappa above something like .75 in order to achieve a lower limit above something .61. And so the sponsor was recognizing -- they may not necessawrily reproduce the .85, but they expected, based on the 206 study, to get a fairly high kappa.

DR. NEATON: Right.

DR. KOCH: What they didn't recognize was that when you move from the side-by-side assessment to separate assessments on different occasions by readers, the within component of variability was going to get bigger, not bigger to a problematic extent because that within component of variance is important to their confidence interval. The confidence interval does involve the mean. But it does involve that within component of variance as well.

DR. NEATON: Maybe I could just ask Mark, I mean, so that given what you've seen now and look at, what would you choose as a boundary for kappa as a noninferiority margin?

DR. LEVENSON: I think that's basically a clinical judgment, and I probably wouldn't have an opinion.

DR. NEATON: So that's kind of -- I guess that's where I'm going. I don't think we're there yet, and so that -- I mean, I heard, Mike, you say that basically, the worst kind of thing that can happen is if you -- this is the way I interpreted it; correct me if I'm wrong, is that I do a stress test. And I look at if there's a difference, maybe, or either score alone. And I send somebody to the cath lab totally unnecessarily; or I do a stress test and miss something really important and don't send them to the cath lab.

DR. HARRINGTON: Okay. So this is going to be the essence of our discussion this afternoon. And there's still two people who want to ask questions.

So if it's okay, Jim, that is the essence of the discussion this afternoon. So let's go to Frank, then Mori, and then we're going to break for lunch

DR. TATUM: I have a question.

DR. HARRINGTON: Is it related to what we've just talked about?

DR. TATUM: Yeah. Let me --

DR. HARRINGTON: Let me put you, then, after Frank and Mori, and who have been waiting.

DR. BENGEL: I have more of a more general comment and just a small question to the sponsor.

I think we've discussed a lot about our confusion with what kind of endpoints to use, and we've discussed a lot about statistics, should we use outcome as an endpoint, should we use coronary angiography as an endpoint, should we use coronary angiography as an endpoint, or should we maybe use noninferiority to an alternative approach as an endpoint?

But I think we have a pretty good endpoint that was used in Phase 2 of this study, and that is quantitative flow measurements, the flow wires. And these results, I think, showed relatively nicely that the agent binodenoson resides in a flow increase or in a degree of vasodilation that is very much within the range of adenosine.

That's where my question comes in. I'd like the sponsor to explore a little bit further on that because we've only seen group data. I'd like to see correlations. Was there good correlation between the flow increase, between adenosine and binodenoson, in subjects on an individual basis?

Because if that's the case and both agents result in a similar amount of vasodilation, then this says to me that all the discussion that we've had about statistics are probably an issue of SPECT methodology, of myocardial perfusion imaging, rather than the agent that we're discussing.

In other words, if that really was the case, that both agents result in a similar amount of vasodilation, I would think that it is probably in some way justifiable to simplify the approach of analysis in the Phase 3 trials just because I also think that the sponsor has chosen a pretty ambitious approach of analyzing the Phase 3 studies.

They use the continuous scale there, as compared, for example, to regadenoson, where it was more the number of segments; they tried to use a summed difference score here. So if the degree of vasodilation of both agents is comparable, then I would think that the Phase 3 studies should be discussed in a different way.

DR. HARRINGTON: So let me -- we're going to take -- that's a very interesting question because it moves some things in a different direction.

Dr. Udelson or Dr. Carter, do you actually have data on the hyperemic response or the flow response? You don't have to show it to us now. Maybe we can tee it up after lunch. I know that what Dr. Udelson showed us was a single patient taken from an AJC article unrelated, necessarily, to these studies.

DR. CARTER: So two points. First of all, we'll look to see how much data we actually have, and we'll be happy to show this to you. We have a little bit of data from the study that preceded 206, Study 202, and I'm going to ask Dr. Udelson to come up and talk to this because I think this is relevant.

The other point to make is that actually we did discuss at an early stage using coronary blood flow as the marker of efficacy in our clinical studies, and this was declined by FDA.

DR. HARRINGTON: Yes. That was going to be the second part of my question, was for Dr. Rieves to comment on whether a hyperemic response is enough in an agent whose goal is to induce a hyperemic response. But we can have, maybe, Dr. Rieves answer that in a moment.

Go ahead, Dr. Udelson.

DR. UDELSON: Put this slide up.

Dr. Bengel, we don't have correlation data to show you from the 202 study. I'll just reiterate what I showed as part of the core presentation. The patient example, as Dr. Harrington mentioned, was one patient from this study, just an example for those not familiar how these studies are done.

But I highlighted here the dose that went on to the Phase 3 trials. And this is coronary blood flow velocity reserves on an index of coronary hyperemia. Percent of the coronary blood flow. Velocity reserved, compared to the referent intracoronary adenosine, was almost 100 percent, a wide range but, you know, when you look at another way of measuring this similar to adenosine, which itself has a wide range.

I'll also mention -- I know some of you are very familiar with this concept -- that you don't need to get 100 percent for SPECT tracers because they're actually not very good blood flow tracers. There's a rolloff phenomenon, and SPECT, thallium, sestamibi, cardiolyte, tetrofosmin cannot track in the 90 percent flow range. So this looks pretty good. So we do not have the individual patient correlations for you at the moment.

DR. HARRINGTON: Dr. Rieves, do you want to comment on, or Dr. Unger, on the suitability of inducing a hyperemic response as a regulatory endpoint?

DR. UNGER: Well, I don't think it's a regulatory endpoint. But, I mean, one could kind of question whether these are actually imaging agents. These are adjuncts to imaging, really. And a lot of the variability that you see with adenosine, you don't know how much of that is the adenosine versus the imaging modality, obviously.

But I had a question for Dr. Udelson. In terms of the coronary vasodilatory properties of the drug, the way this is done experimentally is you do a 10- or 15-second coronary occlusion, and you look at reactive hyperemia, and you look to abolish reactive hyperemia with your coronary vasodilator. That's how you do it in a dog lab, for example. It's done clinically with an angioplasty balloon.

So I wonder if you had any data on that, that you've obtained during the development program from a cath lab where you did transient balloon occlusion, looked for a reactive hyperemic response.

DR. UDELSON: No. We do not, Ellis. And I don't think no animal data along those lines, either. It is challenging in humans, but doable, as you know.

DR. UNGER: I mean, for what it's worth, I'm not a nuclear person, but I do have a lot of experience in animal work. And in dogs, adenosine is not a particularly good coronary vasodilator. There's an old German drug called Chromonar that's fabulous. And adenosine isn't all that good.

DR. HARRINGTON: Before we break for lunch, I'm going to let Dr. Krantz and Dr. Tatum ask a quick question.

DR. KRANTZ: I'll be really brief. Sebastian mentioned earlier that weighted kappa is not really the best test. And I think Dr. LaVange mentioned using the intraclass correlation coefficient. I just wanted to bring that up. She didn't show us any data about the ICCs. And I wonder, is that something that we should be considering across a spectrum of end points?

DR. HARRINGTON: Dr. Levenson or Dr. Rieves, do you want to comment on that?

DR. LEVENSON: I haven't actually now thought of using that statistic or this purpose, so I actually don't have anything to say about that. Maybe the sponsor does.

DR. HARRINGTON: So why don't we mull on

that one over lunch. And we'll come back to it.

So let's go to Dr. Tatum.

DR. TATUM: Yes. The question came up on the analysis of the safety data. And I think I read something that we weren't really looking at safety data at this meeting because it was still being analyzed. Is that correct?

DR. RIEVES: Well, all our analyses are ongoing. But in terms of challenges, the question on why are we having the committee, the safety data appear readily interpretable. They actually look very straightforward. We don't really need all that much assistance evaluating that. But these efficacy data in particular is what we're hoping to focus on.

DR. TATUM: Well, we've seen some numerical data, but we've seen no statistics on the safety.

DR. RIEVES: Right. Right. Again, we brought one question to the committee.

Addressing the statistical aspects for the safety concerns, we're almost teetering into labeling, if you will. And it does come into play. But that somewhat comes at the point that we look towards actually approving the product and working out labeling.

DR. HARRINGTON: Well, Dr. Rieves, let me see if I'm getting at what Dr. Tatum's question is, which is that if we're being asked to consider, particularly if something is equivalent or no worse than another, part of that balance is that you might give up a little bit if its safer.

Is that the essence of your question?

DR. TATUM: Correct.

DR. HARRINGTON: So it might be something that we want to have some -- that would be the classic noninferiority discussion.

DR. RIEVES: Yes. We're fine with discussing that.

DR. HARRINGTON: So why don't we break for lunch. It's now 5 past 12:00, so let's come back at 5 past 1:00, and we'll get started with either the public hearing or more questions to the sponsor and FDA.

(Whereupon, at 12:05, a lunch recess was taken.)

A F T E R N O O N S E S S I O N

DR. HARRINGTON: Why don't we go ahead and get started?

First off, there are no open public hearing speakers, so we're going to move on to both discussion and our ability to ask more questions before we move into the official questions that the FDA have asked us to consider.

The sponsor told me just after lunch that they have answers to some of the questions that you all raised in the late morning. So let's do that. And then I had a series of people who were waiting for their sponsor questioning, starting with Darren, moving to Sebastian, going to Jim Tate and John Flack. Neil, I'll add you to that. And then we'll continue to open it up as people want.

I would like to try to do all of the questioning in the next hour, hour and a half, so that we can then spend a lot of time discussing. But we can certainly play it by ear.

So Dr. Carter?

DR. CARTER: Thank you very much,

Mr. Chairman.

So the first question that we were asked to address related to the intraclass correlation data, and had we done those calculations. And the answer is that yes, we have. And I'll ask Dr. LaVange to come and address that.

DR. LaVANGE: So in my presentation, I introduced the intraclass correlation coefficient really as a means to better explain how kappa works and why kappa might not have been the best measure for the study design that we had.

So if you'd put the slide up.

The intraclass correlation coefficient looks similar to the kappa in value, which is not surprising. The adenosine/adenosine arm, for reference, in 305 was .64. The binodenoson/adenosine arms in the three studies range from .41 to .58. And if you would, if I could have core slide 54.

So the intraclass correlation coefficient, if you could bring that up, was put up here for two reasons. One, it avoided the categorization that we felt was hurting in kappa because, as I illustrated, you could have full agreement on the diagonal and still be off by 2 or 3. Some difference units, you could be off the diagonal and only disagree by 1 if you were on the border of the categories. And so moving to the ICC gets rid of that issue.

It has a similar issue to the kappa in that you're bounded by how big this can be if your population is skewed towards normal. You just don't have enough patient-to-patient heterogeneity in your population for the total variance to be big relative to the numerator, which is what drives ICC to have higher numbers. And we feel like the kappa and the ICC on the adenosine/adenosine arm are a pretty good bound for where we can get with binodenoson to adenosine.

Now, with the kappa, it's true that a high kappa means you've got really strong underlying correlation. But the converse isn't necessarily true. A low kappa doesn't necessarily mean you don't have the strong correlation. And that's pretty well accepted statistically, and Dr. Koch can talk more about that.

We went through the ICC, and then because of the fact that the patient population was skewed so heavily to the normals and the milds, which we felt like put a ceiling on the kappa, we were just not going to get above the .5, .6 range, which in fact the adenosine/adenosine comparison confirmed.

We wanted to focus on the numerator, which is the test/retest within-patient variability. And we did that with our revised primary analysis. And while it's true that those mean paired differences are in fact equal to the difference in the means of the two agents, I don't agree exactly with what Dr. Levenson said because you can have a zero mean and have values discordant in either direction cancelling each other out. But the confidence interval test would likely fail because you have variability.

If you could go to the next slide.

The confidence interval is a function of that sigma hat w, which is the patient-to-patient variability.

So if I had a lot of discordant pairs, one in one direction and one in the other, cancelling each other out, got a zero mean on my primary endpoint, my confidence interval, which is my test, would probably fail because I would be outside the bounds. And that is also confirmed by the results that we had on the absolute differences, which can't cancel each other out because they are all greater than zero.

And if you looked back at the absolute differences, which I believe -- I don't know the core slide -- 81 -- I'm taking you through this quickly because you've seen it before. But the absolute differences, which are not able to cancel each other out, in the 305 study, the difference between the mean, absolute differences for binodenoson/adenosine and adenosine/adenosine has a confidence interval which is pretty tight, and it's around zero.

So I think that allays the concern that we had a funny primary endpoint where discordant values could cancel out and everything would look good. I don't think that that could happen, and I think the data shows that it didn't happen.

Then I'll ask Dr. Koch, there's maybe some other things to add about kappa and ICC quickly for this question.

DR. KOCH: Well, this comment mainly applies to Dr. Levenson's confidence interval on the difference between the kappas for the patients in 305 who were receiving both A and B versus those who had received both A and A. And that confidence interval, as he noted, is wide. And it's also using kappas, which we don't think is very informative.

But we did look at a confidence interval on the ratio of the within-patient variances that applied to the BA sequences versus the AA sequences. And if that confidence interval can come up, that would be helpful. If it's not able to come up --

Okay, this is fine. So we'll go ahead and put the slide up.

We looked at it as square roots because these are within-patient standard deviations. And for Study 305, the ratio of the within-patient method-to-method standard deviations is .87 to 1.15.

As an exercise, we also did similar intervals relative to 302 and 301, although they do not have their own adenosine/adenosine arms. But you can see these confidence intervals on this as a measure within patient variability. The closeness of the two determinations from the two methods are fairly precise.

DR. HARRINGTON: Thank you.

Dr. Levenson, do you want to comment or add to that? And then maybe I'll ask Jim Neaton to help us out.

DR. LEVENSON: No. I have no comments.

DR. HARRINGTON: Jim, any comments or questions?

DR. NEATON: I mean, the intraclass correlation, I take it, was for the SDS. So, I mean, you're right. The intraclass correlation by that difference can be written as the between-subject divided by the total.

And so the between-subject variability, you choose a homogeneous population, then it's going to be smaller relative to the total, but it also can be smaller because of the within-subject variability. And both of those are operating here.

So I don't think that it really adds that much. I think your question, at least in my mind -- there's still an issue about whether that's a difference, which is, is it clinically relevant or not that we have to come back to. But I guess I'd be interested in seeing, rather than the difference of differences, which I think just complicates this unnecessarily, just the stress score.

DR. KOCH: Yes. I understand your question. There is the statistical literature that says intraclass correlation will behave relatively similarly to weighted kappa. And so if you were to look at the stress score, you would probably see that the confidence interval on the intraclass correlation would look basically like the confidence interval that was shown previously on the kappa.

But basically, the perspective here was that the intraclass correlation was simply a vehicle to recognize that what really needed to be targeted was the within-patient method-to-method variance. And what one wanted to try to emphasize was that that colony was small in its own right, aside from dilemmas in intraclass correlation or kappa. And that was what we were trying to communicate in this most recent set of results.

If your within-patient variance is small -- that's the method-to-method variance -- then the two methods will be tracking one another relatively closely.

DR. NEATON: That part, I agree. I mean, I think the data -- what you have in terms of the adenosine/adenosine comparison as well as the AB comparison, it appears the standard deviations of the differences that we saw were very similar. But we're back to, I think, the square about whether that's the right comparator.

DR. HARRINGTON: Yes. And we're going to come back to that.

Dr. Carter, did you have other --

DR. CARTER: Yes. So quickly, we were asked whether or not we had any odds ratio data. We don't because we didn't do the epidemiological work or what have you that were required for us to generate or to be able to express these.

We were asked if we had data -- or what

proportion of the patient population in Phase 3 had a history of diabetes. Approximately a third in all three studies, and they were randomized subsequent to the history having been identified. And we did not do any -- at least we don't have access today to data that would do a subset analysis relative to diabetes and we don't have any injection fraction information to give you.

DR. HARRINGTON: When you say you don't have injection fraction data, did you not measure it as an entry point or did you not get it during the course of -- because it's frequently gotten, as you know, during stress testing.

DR. CARTER: Yes. Jim?

DR. UDELSON: I think what we have is a history or left ventricular dysfunction or not, which we can eventually get for you.

From the core lab analysis -- and of course, you're correct that gated SPECT imaging is done. And that was used to help the readers differentiate infarct from artifact. But I don't believe we captured fully the ejection fraction information.

DR. HARRINGTON: And along those same lines, Dr. Udelson, we'd asked the question about data that you might have had on LV dilatation or lung uptake, other things that people might be concerned about if the two tests were not picking up that same group of patients.

DR. UDELSON: Right. Lung uptake actually is most useful in exercise as opposed to pharmacologic stress, where the demand, of course, is very different. So we did not capture that, and it's actually never been -- it's not as useful during technician studies.

Transient dilatation I don't believe we captured. But I think if we had, the prevalence would have been pretty low in this population.

DR. HARRINGTON: Other things, Dr. Carter?

DR. CARTER: Thank you. We were asked whether or not we had any statistics on the safety and tolerability data, and indeed we do. Dr. Udelson did show these data. We have an extensive set of slides that we can go to, but I'm sure that he can just summarize for you.

We focused, as you remember -- as a prespecified objective of this whole development program, we design new studies to allow us not just to show concordance or agreements in terms of efficacy measures, but also to allow us to collect in a prespecified way and to compare the safety, adverse events, tolerability, patient preference, and so on between binodenoson and adenosine.

DR. HARRINGTON: Yes. And I think this is an important part of the discussion because although Dr. Rieves is correct, they're not asking us to comment on the safety issues per se, I also think Dr. Tatum's correct in that one has to consider the safety part of the equation in determining how much one might be willing to give up or what level of uncertainty one might be willing to accept with the comparison of the two agents.

So I think this will be a really important part of the discussion and people should weigh in.

Dr. Levenson?

DR. UDELSON: Thanks. You know, in all of the slides -- we show you a million slides that we've been looking at for months -- important points do get lost sometimes. I think a couple of people mentioned the statistical analysis of the side effect data. And I think it is important because right from the beginning, the whole idea of this is a selective agent that has fewer side effects. So there was a lot of thought that went into this.

Can I have the slide up, please?

So just to reiterate from the core presentation, it was prospectively defined in the protocols that the analysis of side effects would be very rigorous. There was a sequence to it to account for multiplicity.

The sequencing was based on the previous studies, in particular 206. So statistical testing was done on each item in the sequence. And when an item did not reach statistical significance, no further significance testing was performed throughout the sequence.

Can I have the next slide that I have here?

So this is actually a more in-depth analysis than I showed you this morning. So this is the entirety of the sequence. So part 1 of the sequence was the incidence of second- or third-degree AV block.

You know what? Could I have that one back up here? Thanks. And this is in Study 305.

So the incidence of second-or third-degree AV block was statistically significant in favor of binodenoson. Then you move on to the next item in the sequence, which would be overall symptom bother.

The difference in proportions, favoring more patients being not at all or a little bothered versus some or a lot bothered over here. It was highly significantly different. Patient preference, highly favored binodenoson. You move on in the sequence.

Flushing -- and the actual sequencing was both the incidence and the intensity of the particular side effect. So you see here the incidence of flushing was lower with binodenoson; significant. You move on to the intensity. That was lower.

Move on to the next part of the sequence.

Can I have the next slide? Slide up, please. Chest pain was next in the sequence, and I won't belabor this, but incidence and intensity, shortness of breath; incidence and intensity. And then in this particular trial on 305, nausea was lower. Incidence and intensity. Headache was not, so the sequence stops there.

A similar pattern --

We just showed the 302 data? Okay. But I'll go through this much faster.

Here's the 302 data. Same pattern. Second- or third-degree AV block. Patient bother. Patient preference. Flushing. Incidence and intensity. All highly significant.

Next slide, please. Thank you.

Chest pain and dyspnea. Incidence and intensity highly different between the two, less with binodenoson; on nausea, not quite, so the sequencing stops there at that point. And I'm not sure if we have it, but not to belabor the point, almost the same was seen in Study 301. So all three of the trials, the side effect data, we thought, were, A, rigorously planned and analyzed and showed the significant data you've seen.

DR. HARRINGTON: Dr. Udelson, can I clarify two things? Could you define second- and third-degree AV block for the purposes of this study?

In other words, if you had three beats of AV block, did that qualify? Or did they have to be something that was prolonged for some period of time, or perhaps even account for some symptoms? Because, as you know, if it's very transient, it may not matter.

DR. UDELSON: Yes. It often is very transient. This was investigator-reported second- or third-degree AV block.

DR. HARRINGTON: And a similar question in terms of whether or not we should think it matters, this question about asking patients which they preferred, you noted that the site was unblinded to the order of the scans. I understand the patient was blinded. But was the person who asked the question blinded or unblended, or did they already have knowledge of what scan happened and in what order?

DR. UDELSON: Dr Barrett, who was involved in the trials and the operations, will answer that.

DR. BARRETT: The site personnel was

supposed to be -- remained blinded all the way through. Only the investigator was supposed to be unblinded as to the nature of which scan was which so that he could refer. But we don't have any definitive information as to whether a nurse or a study coordinator might have been unblinded in some cases.

DR. HARRINGTON: So the intention was to keep them blinded, but there's the possibility that the questioner might have been unblinded.

DR. CARTER: And of course, we were concerned about this. So we actually put in place a very rigorous independent audit process to look at every possible step and every possible eventuality to assure ourselves that unblinding actually had not occurred. So the audit confirmed, as much as the audit could tell, that there was no unblinding.

DR. UDELSON: The bother question, how much did this bother you, was asked after each individual study, before --

DR. HARRINGTON: I see. Okay. Before knowledge would have been possible?

DR. UDELSON: That's right. And, in fact, we have -- in terms of auditing to make sure, there were sort of time stamps reviewed by the monitors to make sure that the release of the blind to the PI, so they could know which was the adenosine, was done after the bother question answer had been recorded on the case report forms.

DR. HARRINGTON: So would it be fair to say that the bother question might be more rigorous than the preference?

DR. UDELSON: That would be fair to say.

DR. HARRINGTON: Dr. Tatum, did you have a comment that you wanted to make on this?

DR. TATUM: There was another study on bronchospasm, which was a separate study, since I think that's the most important issue entirely.

DR. CARTER: Dr Barrett, will just give you a brief rundown on that.

DR. BARRETT: Yes. We did conduct a study on patients with mild intermittent asthma in order to determine whether or not doses of binodenoson did produce any bronchoconstriction, which we defined as a decrease in FAV of greater than 20 percent. This was done at seven different sites by pulmonologists or allergists using standard pulmonary function tests, and we didn't see any evidence of any bronchoconstriction in any patient tested, any patient who received binodenoson. Now, of course, these patients did not receive adenosine as a control because it is contraindicated in these patients.

DR. HARRINGTON: Dr. Tatum, do you have a comment on that data?

DR. TATUM: Well, since is the biggest problem I think I've had to deal with in this area, I think this is very important to the whole safety issue. It would have been nice to have a control, but I understand why they didn't do it.

DR. HARRINGTON: So this matters to you?

DR. TATUM: Yes.

DR. HARRINGTON: Okay.

Mori, I'm going to add you to the list.

Dr. Levenson, do you want to comment on the statistics of the safety?

DR. LEVENSON: Well, no, not on the safety.

DR. HARRINGTON: On the other? You can go ahead.

DR. LEVENSON: Okay. Just a quick response to the confidence interval and the difference that the variation would address, like values way away from zero. The confidence interval is actually a confidence interval on the mean, so as the sample size gets smaller, that will get smaller as well. So it's not at all a measure of the spread of the distribution.

DR. HARRINGTON: So you're not convinced that this alternative method is one with which we could --

DR. LEVENSON: I'm not convinced this confidence interval will protect you against symmetric values away from zero because as the sample size gets larger, you can have values away from zero, but the confidence interval of this mean difference will shrink to zero.

DR. HARRINGTON: Dr. Koch?

DR. KOCH: Yeah. This is Gary Koch. Well, again, the tendency for there to be positive or negative values in either particular direction will indeed move the mean towards zero as the sample size gets bigger and bigger. But the more variable those differences are, the within-patient difference, the variance of that within-patient difference, will be correspondingly bigger.

So you need to have the distribution fairly concentrated about zero in order to get the variance small. So that was essentially the way in which the method was working. It had to have both a small, within-patient variance as well as a mean near zero.

Now, to further address the robustness, that was the role of CC-81, where we actually, based on motivation from the FDA -- slide up, please -- focused on the mean of absolute values. Absolute values are positive. And we certainly agreed with the FDA that this was an important graph to look at because it shows the cumulative distribution of absolute values, which are always positive. And what we then did was to apply the methodology that the sponsor had, but now we have to compare the binodenoson minus adenosine versus the AA sequence.

So we have to do a two-sample comparison.

We have to get the mean of the absolute differences for BA and compare that with the mean of the absolute differences from AA to basically show that those patterns of curves are basically the same.

Now, this was a follow-up analysis motivated by the FDA analysis that called our attention to this cumulative distribution. But here there are no positive or negative values. They're all positive. But we have to have the AA group to compare against.

DR. LEVENSON: Well, my first comment, would you agree that as the sample size got larger for the same within-patient variation, the confidence interval on the difference would get smaller?

DR. KOCH: Yes. That is correct. I think we had a slide that showed that. But for any particular sample size, you need the within-patient variance small.

DR. LEVENSON: Okay. But it's really not a measure of the spread of the distribution and all; it's the confidence interval and the mean.

DR. KOCH: The confidence interval that was used by the sponsor is comparable to what is used in pharmacokinetic studies to basically show that pharmacokinetic parameters like AUC, it is a population equivalence, but the methodology is essentially the same as what was used in pharmacokinetic studies.

Now, when we work with the mean of the absolute values, that's a robustness assessment that goes further. And we agree with the dilemma that you've identified in the original proposed method. That's why we thought these other results were relevant.

DR. LEVENSON: I just have a quick question, if you could bring back the slide on the cumulative distributions.

So these confidence intervals, are these confidence intervals on the difference in the cumulative distribution functions or are these sort of confidence intervals on the patient level differences? I'm not sure I'm making that clear.

DR. KOCH: Well --

DR. LEVENSON: I mean, you have a curve for binodenoson and a curve for adenosine here. Is this a confidence interval on the difference of the curves or what is it?

DR. KOCH: The confidence interval addresses the mean of the absolute difference. So there's an SDS score for B and an SDS score for A. We take that difference and form its absolute value. So we're working with the mean of the absolute values for B versus A compared to --

DR. LEVENSON: On a patient level?

DR. KOCH: Yes. On a patient level. So we have an absolute difference of B versus A at the individual patient level, the same thing that the FDA looked at when they produced the display that corresponded to binodenoson versus adenosine.

We have an absolute difference, SDS for B minus SDS for A on each patient. Take the absolute value. Calculate the mean of those absolute values. Do the same thing for A versus A, and then have a two-sample confidence interval on the difference.

That is a fairly precise confidence interval, and it doesn't have the dilemma of positive values cancelling negative values.

Now, equality of distribution is consistent with equality of means. If you had similar means, you would expect to have similar distributions. So it's a comparison of the distributions through the corresponding means.

Now, CC-80 looks more fully at the distributions as a whole, and Dr. Udelson had spoken about this. And then the confidence interval that I referred to previously, which was essentially the confidence interval on the within-patient variances, where we got the within-patient variance for the BA sequence against the within-patient variance for the AA sequence, that's working with averages of squares. And a square gets the distance between two determinations on the same patient. So the average of squares for B versus A is comparable to the average of squares for A versus A.

DR. LEVENSON: Well, I would agree that some of these additional analyses get a compatibility within patient, but I still feel that the revised efficacy measure is still inadequate.

DR. KOCH: Well, I think that was why the sponsor believed that they would need to supplement their revised proposed method with additional analyses. One of those additional analyses was essentially an upper limit of 10 percent of extreme disagreements, and additional analyses are some of the ones that are being presented here because I think they do recognize that in order to support their primary method, they need other results that say that ways in which it could have gone wrong, it didn't go wrong.

DR. LEVENSON: Yes. But as a primary outcome for a confirmatory trial, I think it was deficient. It should have done more by itself. And that 10 percent criteria only protects against extreme discordance, which may --

DR. KOCH: Well, I think the sponsor there was simply reacting to the notion of totality of the data. And they had a primary criterion which, if it failed, the study would fail; whereas if it were successful, they recognized they had more work to do, and they did indeed try to do more work to provide assurance that success on that criterion was indeed a reasonable basis for success.

DR. LEVENSON: Yes. I'll just say again that for a primary outcome for a confirmatory trial, I would expect it to show more that this measure would.

DR. HARRINGTON: So I'm going to go to Dr. Halperin, Neaton, and then Unger.

DR. HALPERIN: Just a very basic question that gets to the issues that Dr. Levenson was raising and to the fundamental regarding the equivalence or noninferiority of the new compound to adenosine. And that is, if the sponsor could comment on the method used to derive the sample size for these Phase 3 trials.

DR. CARTER: Dr. LaVange?

DR. LaVANGE: So the sample size for Study 301 and 302 was based on the kappa threshold of 61. And the sample size calculation, which I thought we had a backup slide on, so I may need my notes, gave us 90 percent power to exceed the threshold of .61 because when 301 and 302 were designed, then that was the primary analysis. It was, in fact, overpowered for what eventually ended up being the revised primary analysis.

I don't know if that helps.

DR. HALPERIN: So a sample size determination with the original design?

DR. LaVANGE: It gave us 90 percent power for a kappa to exceed .61 with significance at 05, which meant the lower bound of the -- I mean, the upper bound of the confidence interval would be -- the lower bound of the confidence interval will be to the right of .61.

DR. HALPERIN: And what was that in?

DR. LaVANGE: 376, or -- 320? I don't have the notes, 320-something.

DR. HARRINGTON: Does that answer your question Dr. Halperin?

Dr. Neaton and then Dr. Unger.

DR. NEATON: I'd like to kind of ask two questions just to follow up on that one, if you don't mind, too. But maybe you could just go back to the other issue first that Dr. Levenson brought up.

I think Dr. Koch had said this in his response, but I guess one thing that I found some assurance with in looking at is on page 56, the fact that the within-person -- the standard deviation of the differences, is what this is, for the adenosine/adenosine comparison was very comparable for the BA comparisons in all three studies.

So you're right. You have to look at more than the difference. And that's very important because that difference can be zero. And actually, I've been burned in studies before where it's zero, and without a focus on the standard deviation, you just would totally have made the wrong judgment about equivalence between two items. But if we're willing to calibrate the standard deviation that we see and, therefore, confidence interval around it by what you observed with adenosine/adenosine -- although realizing it's limited in this set of studies; there's only one trial that did it -- that gives me some reassurance.

I also had a question about the sample size. So you redesigned the studies but you didn't readjust sample size. You left the sample size the way it was. I had a similar question.

Also, why did it take two and a half years to figure out what to do in terms of the redesign?

DR. CARTER: To answer the last part of the question, two and a half years, yes, in the grand scheme of this particular program, that doesn't seem such a long time. But there was obviously a lot of interrogation going on. There was a lot of-back-and-forth discussions, both internally and with the agency. And bear in mind that the idea was initially held in the cardio-renal division before it moved into the medical imaging division.

So there was a fair amount of inefficiency, let me say, in terms of being able to get to where we are today.

Relative to your first question, which is on the sample size calculation, perhaps I can ask Dr. Koch to come back up?

DR. NEATON: I understand it. Neither of the last two studies, the sample size was modified even though you changed the end point?

DR. KOCH: That's correct. The planned sample size that those studies originally had as they might have targeted kappa was more than ample to provide 95 percent power, or better, for what was the new criterion. And that simply comes from the fact that you have more power to address issues on differences in means than on something that is like a correlation coefficient, which is what the kappa statistic is.

DR. NEATON: I kind of appreciate that, although it does raise the question, did you just back into this end point because you thought you had the sample size to investigate it as opposed to kind of develop it kind of with a thoughtful approach in terms of the clinical relevance issues that we've been kind of trying to grapple with.

DR. CARTER: Well, no. We absolutely developed it through careful consideration and through the determination of what we believe to be a robust clinical analysis approach. So this was not a backing into at all. It was entirely as prospective --

DR. NEATON: So maybe the question I haven't heard is the one I asked this morning, is your justification based on clinical relevance with the other paper on page 51 and what a difference of the magnitude that half a standard deviation of the stress score would mean in terms of predicting clinical outcomes. I mean, that must be available somewhere in the literature for you to gauge.

DR. KOCH: I'll just make a brief comment. The reasoning -- "supporting the revised method" is the reasoning you're already heard, to focus on a confidence interval that would capture the information in the within-patient variance.

That method was recognized, when the sample-size calculation was done, to have better power than what the original method had. So the sample size did not need to be adjusted for that reason.

The margin that was set was set on the basis of clinical reasoning, which Dr. Udelson can speak to further, together with being half of a standard deviation, which had a statistical reasoning.

It was recognized that the confidence intervals probably needed to perform better than what the margin was. So that was why Dr. LaVange noted that, ideally, not only would the intervals be internal to minus one and a half to plus one and a half, they'd be internal to minus one and a quarter to plus one and a quarter, and perhaps would even be internal to minus 1 to plus 1. And as you tighten the margin, then, of course, this notion of higher power starts going down towards usual power.

But again, I think Dr. Udelson should comment on the clinical relevance of the one and a half margin.

DR. UDELSON: Thanks. So from the clinical perspective, the margin was in part based on what difference in summed difference score would be clinically relevant. And as we started to think about this, it's a little problematic because it's a continuous scale. If you have thousands of patients, which some prognosis studies do, you see a continuous increase in event rate.

So we tried to find studies where we could pull out something that would support a small -- because we wanted this difference to be small so that the equivalence would be robust and really pass the straight face test.

Three of the papers that are referenced I think on page 51 of the briefing document, it's fairly straightforward. In one study of about 7,000 patients, greater than 5 percent of the myocardium that's ischemic is associated with an increment, a statistically significant increase in the prediction of cardiac death. And I believe it's 2.9 vs. .5 percent, something like that, from below 5 percent to above 5 percent. So 5 percent of the myocardium corresponds -- 5 percent ischemic corresponds in this system to an SDS of 3.

The first paragraph on that page you referred to, I agree with you completely, is somewhat convoluted. And as I read it again at lunch, I completely agree. The bottom line of that is that the numbers came from consultation with the author of that paper because it's hard -- if you just have the paper to pull out from the adenosine prognostic score on paper, it's hard to find your way through those data.

But we, in consultation with the author of that paper, using a prognostic score with adenosine based on about 5,000 patients and outcomes, it was thought that an SDS of 3 would correspond to an increment in event rate that would be clinically relevant.

DR. NEATON: Do you feel that using that kind of data is the best way to assess the clinical relevance as opposed to the issue that Dr. Domanski brought up earlier in terms of predicting going to the cath lab accurately?

DR. UDELSON: Well, that's a very fundamental question and it'll get me back, I guess, to answering Dr. Domanski's point earlier. The scores, the 17 segment model scores, were actually not designed to be parsed at a particular place for sensitivity and specificity analyses originally, but they originally grew out of large prognostic databases and they have a fairly continuous relationship with outcome risk.

Someone -- I can't remember exactly who it was this morning -- noted from one of the slides about the very modest positive predictive value for outcomes. In other words, if you have a severely abnormal scan, the event rate was, let's say, 20 percent, meaning 80 percent of the time it's wrong.

But that's event risk. I mean, you parse people into risk groups -- low, medium, and high -- but many people with a very positive scan do fine because in contemporary practice, they may have 3-vessel disease and do fine on medical therapy, as we've seen in contemporary trials, like COURAGE and others.

Nonetheless, the imaging data parses them into risk categories that, for publication purposes, are statistically significant. And clinicians respond to those data by saying, if I have a patient with a severely abnormal scan, they are a higher risk of an outcome event. And if I cath them, and if I revascularize them, I believe I am lowering that risk.

But that last piece of that sentence is actually, scientifically speaking, not so supported, even though all of us who practice cardiology do that every day. And we could support it from propensity-matched analyses of trials. But at that exact point, do we really know that from randomized trials? No. But that's how we practice.

So the extent of ischemia drives clinical decisions. And I think to step back in the big picture of this, the general idea was if the extent of ischemia is fairly similar between new agent B and old agent A to some relatively narrow range, which we tried to define as best we could, then we take the step. And I take your point about the surrogate of a surrogate. But we make the leap that similar clinical decisions would be made.

DR. HARRINGTON: So, Sanjay, is it on this exact point? Because otherwise it's Dr. Unger.

DR. KAUL: Yes.

DR. HARRINGTON: Okay.

DR. KAUL: I just want to share a statement in that paper from Rory Hachamovitch, that you're referring to, which sort of illustrates the caveats of using this estimate to estimate the risk.

"Predicting risk based solely on the relationship between myocardial perfusion defect and outcomes would result in a mis-estimate of risk." And for the reasons that people have already gone over, you know, the diabetics, the elderly, the patients with LV systolic dysfunction.

So to estimate the clinical relevance based on this unstable estimate is a slippery slope.

DR. UDELSON: No. I agree in concept, of course, with what you're saying in practice. And in fact, you know, the strength of that prognostic score paper, and it is an extremely important paper, was that it, for the first time, incorporated clinical data, responsive, heart rate, EKG -- you know, that is how we think, really, in real life.

You know, when I'm reading a scan, I have all those data, and I love that paper because it really -- it wasn't just the images. But I guess you could say in a regulatory environment of a clinical trial like this, where the readers are sort of locked in a room with an image and score these segments, and then we are trying to figure out is this image similar to that image, we don't incorporate the clinical data into the reading.

Now, I'm not saying that's right or wrong. You know, ideally, perhaps the best way to do this is to give a reader all the information and see what decision they might make, theoretically, and then give them all of this and see if they might make the same decision, but there's problems with that as well in a clinical trial.

So I completely agree with your point that the clinical data plus the imaging influence the outcome. A normal scan in a young person is associated with an event risk of less than .5 percent. A normal scan in an 80-year-old diabetic woman is associated with an event risk of 3 percent because the pretest probability influences the post-test risk, again, as you have written about.

So we did not incorporate that concept into the analysis because of the particular constraints in reading imaging in trials.

DR. HARRINGTON: Dr. Unger?

DR. UNGER: Thanks. This is a double question, I guess.

Could somebody put up slide CC-52? The first question is kind of a double question in itself for the applicant, and then I have more of a general statistical question. So this is a totality of the data issue.

There are a couple key differences between this analysis. This was the regadenoson analysis. In one of them we talked about earlier, which this is kind of a three-tier, 3x3 table, whereas the urinalysis was 4x4.

But they did something else. I mean, they preserved some of the spatial information here by counting the number of abnormal segments, 0 to 1, 2 to 4, greater than or equal to 5. So they have some of the spatial information remaining in there.

What you did was you removed all the spatial information. You collapsed 17 regions of interest into a number, a summed difference score, if I understand correctly. So you've more or less thrown away the spatial information.

Then the other thing that was done was you counted scar the same way you'd count normally perfused myocardium because there's no difference with stress. And you showed us the apical segment and it got a score of 4, both at stress and at rest, because it's a scar.

So you basically are saying, we don't care about scar in this analytic plan. We're going to count that the same as we count a normal myocardium. So you've thrown away that information, and some of that information is pretty important.

So my question for you guys is, did you ever analyze the totality of the data; meaning, for each patient, 17 regions of interest, what was the agreement, region by region, for exchange patient? Now you have 17 times as much data as you had before.

So the question is, did you analyze that? And then the other question is, did you ever analyze your data the way the Reg Dennison data were analyzed in the slide that was just up there that's not up there?

DR. UDELSON: Thanks, Dr. Unger.

Can you put this slide back up? Thanks.

Let me comment on this for a second because I actually think the point we're making here is the opposite, to tell you the truth. I didn't see, of course, what they submitted to you. I saw what was in the paper.

The way I understand that this analysis was done is that the readers scored in a 17-segment model exactly as the readers of these trials did today. The data were then collapsed. In other words, if you had a segment, all segments with either a different score of 1, 2, 3, or 4, mild, moderate, or severe ischemia, were called ischemic.

So in other words, all of the information on severity of ischemia was removed. And there's no localization here; this is essentially just how much of the ventricle has any ischemia, without regard to severity. So I would submit that a lot of information was lost.

DR. UNGER: Yeah. No, I agree.

DR. UDELSON: And let's take -- let's look in this middle ground. So 2 to 4 ischemic segments with adenosine, 2 to 4 ischemic segments with regadenoson.

Because the range in any segment could have been a score of 1 to 4, at least theoretically, you could have a study with 2 segments with a different score of 2, so a summed difference score of 4; that would be here. But you could have 4 -- the same patient could have 4 adenosine segments with a score of 4. So that patient could have had a summed difference score of 16 and a regadenoson score of 4 and be counted as agree.

So, in fact, the possibility exists that there's very poor agreement here because they're minimizing the variability, plus there was only one rest study done in the majority of patients. The rest study wasn't repeated, which, as Dr. Neaton has pointed out, adds to the variability here. So, essentially, this is an analysis of the summed stress score.

DR. UNGER: Okay. So your point's well taken. This may not have been the brightest idea, either. But what about analyzing your data, all 17 regions of interest?

DR. UDELSON: I don't think we have transformed it into this.

Can I see the -- we do? Okay.

We also have localization data that I'll show you.

Okay. Slide up, please.

So this is the data from the trials, collapsed into the kind of analysis that was done for regadenoson. And this is the agreement across the three categories, as they did. And you see the numbers here.

DR. LaVANGE: Their analysis was row by row to compute the percent agreement and then take an unweighted average. So each row contributed a third, a third, and a third. And we were able to do the same thing on our data, except we had four categories.

DR. UNGER: Okay. But did you ever do all 17 regions for every patient? That's a lot of data. You could learn a lot from that, maybe.

DR. UDELSON: We have not done that. I think you're also multiplying the variability, I think. It's very granular.

Let me get back to your other point, Dr. Unger, about ignoring the infarction. And I think we addressed that a bit this morning in Dr. Neaton's comments about the summed stress score, which we did propose to FDA to be what we wanted to actually analyze. And we were told no, that is not sufficient, because the summed difference score is the extent of ischemia. And we showed you the summed stress score, which was quite similar to the summed rest score and quite similar to adenosine and adenosine, so even when you account for that.

Can I have this slide up, please?

I think another way to get to your point about localization is here's an analysis that's, as it's typically done in nuclear cardiology studies, correlation with angiography, where these are the different studies, binodenoson/adenosine and adenosine/adenosine sequence, where the correct identification of the region, the vascular territory between LAD and non-LAD -- and I think in the literature it's typical to lump circumflex and write "non-LAD" because there's a lot of overlap and it's hard to do that from imaging, SPECT imaging. As you can see, the percent of exact agreement across here, high 70s, 80s, about the same in the non-LAD. And we have this.

This is for the reader summed difference score.

Do we have the summed stress score also?

And then to get, again, at your other point about incorporating the infarction into the analysis. Slide up, please.

This is now exact agreement by vascular territory using a summed stress score. So 83, 83, 78, adenosine, adenosine, 77. And these numbers are pretty typical of what you see in the literature of the isotopes, et cetera, compared to vascular territory. So this gets a little bit at the localization issue that you were mentioning.

DR. UNGER: Okay. I find that helpful. Could I ask -- do I have enough time to ask him --

DR. HARRINGTON: Absolutely.

DR. UNGER: Okay. Here's the statistical kind of question. I mean, in your garden-variety efficacy study, variability is the enemy. Your effect has to overcome the variability in order for you to win. But this is not that kind of efficacy study. This is a study of agreement. And according to the guidance -- I hate to quote this but I will -- "Both agents consistently give identical results." And

Dwaine can tell you what that means exactly. I know what identical means; I just don't know what consistently means.

But at any rate, when you're trying to show sameness, then the variability is no longer the enemy. The variability is a way that can help you win. And this is the statistical question, which is, I know that if I had dropped both imaging agents on the floor and never injected them, then I would have won. I would have shown agreement. I'm quite sure of that.

So the question is, if you, for example, combine 17 regions of interest, are you obscuring data that in fact works in your favor if you're trying to show agreement? So it's a general question about showing agreement and variability for any of the statisticians.

DR. CARTER: Do you want to take a shot at that, Dr. Koch?

DR. KOCH: Well, since that analysis hasn't been done, it's difficult to assess what its implications would be. What has been emphasized is that within patient variance, the method-to-method variance is sufficiently small for the sponsor's primary method to produce a confidence interval that was successful; and for some of the other methods that we have gone over, to essentially say the method-to-method variability for B versus A is comparable to the method-to-method variability for adenosine versus itself.

Whether we would learn more about that by looking at the individual statements in a similar way, I don't know. I would agree it could be of potential interest. I don't know whether Dr. Udelson has any further comments on this or not.

DR. UNGER: I guess the general question, again, is more about the noise. If you don't do this in an optimal way, then are you biasing the study towards showing agreement?

DR. HARRINGTON: Well, that's the common noninferiority complaint. Right? That the sloppier you do things, the more likelihood you have of showing that one thing is not different from the other.

DR. UNGER: Exactly.

DR. KOCH: Well, again, basically, the

binodenoson in these studies is showing some of the same traits as the adenosine itself. And so adenosine to adenosine, which was randomized in 305 study, is showing essentially the same patterns of method-to-method variability as binodenoson to adenosine. So whatever is involved is involved in the adenosine/adenosine type of thing as well.

DR. NEATON: I wanted to say something similar. I would have been concerned if the average difference was zero, the confidence intervals was in the bounds, but the standard deviation of the difference for the AB comparison was a lot bigger than the AA comparison, for the reason you mentioned.

DR. HARRINGTON: All right. So I'm going to try to get some order here. I know Darren's been waiting. And then I want to go to Neil. I've got John Flack still. I'll add you to the list, Peter. We got Sanjay, Mori. So we'll try to keep some order here. So keep that in mind as you ask your questions.

DR. McGUIRE: So I want to get away a little bit from the technicalities of the statistical handling and get more back to the clinical context here. And one of the things, again going back to collapsing these into 2x2 tables where clinically we typically interpret perfusion studies as positive or negative; and if we do that in Phase 3 studies, we have a range of 26 to 34 percent discordance.

That means there are 26 to 34 percent of patients being reclassified, effectively. And in the epidemiologic world, when we generate statistical models, the reclassification index has emerged as probably the premier analysis tool, testing one strategy versus another. So in the clinical context, I'm concerned when we're reclassifying 26 to 34 percent based on the agent used as the pharmacologic stressor.

In this case, where the reclassification is bidirectional and relatively balanced in all three studies, it works toward the favor of the SDS differences because the net balance is zero.

So I'm still trying to struggle if we can assume -- so the fundamental concern I have is if we interpret the SDS differences in isolation, the fundamental requisite for that interpretation is that the two diagnostic strategies are sufficiently concordant. If they're discordant, then the SDS differences become a little more difficult to interpret because, as we've seen, the bidirectionality may tend to center the result on zero.

So the question I have is, how do we reconcile this apparent discordance? And I understand the challenges of the adenosine test-retest discordance.

How can we at the end of the day take a population parameter and suggest at the end of day that these two diagnostic strategies are similar?

DR. HARRINGTON: Darren, just remind us, what was the discordance in the adenosine/adenosine?

DR. McGUIRE: Thank you. So there's 26 to 34 percent for the binodenoson/adenosine comparison. It's 20 percent for adenosine/adenosine. So numerically, it's pushing twice as much discordance, if you just look at the 2x2 tables. And, again, although the quantitative information across the severities of abnormalities is informative scientifically and experimentally, but in all honesty, clinically we use an all-or-none dichotomy.

DR. UDELSON: Thanks. Let me start with the last point because, actually, I wouldn't agree with that and respectfully disagree.

Imaging, particularly this kind of imaging, is not -- yes, there is a level at which it's normal or not normal. But as I think you might have said earlier today, among the not normal is a range of abnormality, which is actually very important because many, many studies have shown that patients with mild abnormalities, let's say mild ischemia, actually do not need to go to catheterization because they have a low risk outcome which will not be improved by a procedure with some risk, like revascularization.

So as a clinician, you want to sort of restrict -- that may not be the best word -- and again, always in the context of the clinical data -- for the most part, it's the patients on the higher end of the extent of ischemia.

So that concept, that there's a range of abnormality of ischemia that drives clinical decisions was in part fundamental to the structure of the analysis that was set up that preserved, sort of, this degree of abnormality and tried to suggest that there was some concordance between the degree of abnormality one to another. And, nonetheless, when you break it down to normal or abnormal, there's some degree of concordance and some degree of discordance as well.

Now, I think another answer to another part of your question might be the angiographic data because among the discordance, or the 20 to 30 percent that you mentioned, the question comes up, which is right? And so you have to move on to some independent gold standard. And again, we'll get back to Dr. Domanski's point and the points that some of you made.

You know, it is completely correct that the population going on to angiography is a subset. It's not a representative subset. It's a clinically driven subset by the adenosine data. It wasn't the purpose of these studies to create a robust angiographic study.

Nonetheless, there's a lot of data there. And so -- can I have this slide up -- when we ask, what do the -- you know, we spend a lot of time wondering, what do these discordances mean; and is binodenoson inferior because the discordances favor adenosine?

So I think the best we can do is say take the people with discordance in the scores who went on to angiography on the basis of the adenosine data, biased in some way though that may be, and then compare the imaging data to the angiographic data. And without belaboring this, because I showed it earlier, about half the time binodenoson is correct and half the time adenosine is correct.

Now, let me take the opportunity to just move to your right and answer Dr. Domanski's comment from this morning, or address the comment.

In these type of trials, you know, the angiographic data are being read in a core lab, not by the sites, and there's a gross correlation with what the sites think and what a core lab thinks. Usual the percent stenosis is less. And we call a study a true positive if there's an abnormal amount of ischemia, let's say, and a greater than or equal to 50 percent stenosis. And I think Dr. Flack this morning made the point, well, you know, that's pretty kind of old school, and I agree with that. I mean, I don't want to send my patients to the cath lab if they have a 55 percent stenosis; and if the nuclear scan's normal, it's not physiologically significant.

So from the clinical physiologic perspective, that's how we think. But in this kind of regulatory environment, the greater than 50 percent stenosis has been used in the past, and there's history and precedent to it, and, we of course feel obligated to show the data.

I don't think we have it to show you, but we do have data on different degrees of stenosis if you create different cut points -- and, again, this is more for Dr. Domanski's question -- greater than 70 percent, greater that 90 percent, and the sensitivities ands specificities change slightly, as you would expect.

But, again, for the discordances, I think this is probably the best way we can address it by independent standard.

DR. McGUIRE: But again, in the clinical

context, if we are reclassifying 25 to 30 percent of patients, the possibilities are that the experimental agent is superior, adenosine is superior, or it's a wash and it balances out.

And in the absence of a truth standard here, you go to the cath data. And I agree that that's the truth standard. But with this level of discordance in the backdrop, it's my opinion that we may well require a truth standard to prove the utility of this.

DR. UDELSON: You know, and of course, the only thing we have, considering the ROC outcome, you know -- the only thing we have that gets close to, I think, what you're getting at is the 60-day follow-up -- the next one -- is when all of the patients are followed for either death, zero; myocardial infarction I believe was only 6; it's a stable population for a short term; but clinically driven revascularization, driven in part by the adenosine data.

You know, I think what you can take away from this is the binodenoson data, which were theoretically not driving any of the clinical decisions, would theoretically have in the population driven the same decisions.

DR. HARRINGTON: Go ahead.

DR. McGUIRE: So two very quick questions, somewhat related, just trying to get my head around the clinical application of this revised primary end point.

You set up the revised primary end points defining 3 or more SDS difference, 3 or more as clinically relevant, and then as a conservative analysis set up an extreme outlier analysis that is the zero maximum intensity.

But I think if we're going to use 3 as a clinically relevant difference, perhaps we should look at the outliers using a threshold of 3 or more instead of, you know, the extreme. The quadrant boxes at extremes are infrequently populated and clinically rare, I would think. But what would be more clinically relevant is how often was there discordance at a level of 3 or more, as defined as the clinically relevant threshold for the primary endpoint?

Then the second question -- and if you want to -- it looks like we may have a little bit of time to address that -- the second question, just to be thinking, and I'd like the statistical input about this, which is to what degree -- when we're centering the primary endpoint around zero with confidence limits, to what degree do the normal/normal patients influence the outcome?

That is, the majority of patients in all these trials were normal/normal, that is, 0-1, 0-1, influencing a centering of the outcome. So the two questions I would have is if you extract the normal/normal patients, what do the analyses look like, and importantly, what do the histograms look like of the distribution of the SDS deltas without the normal/normal patients concordantly in there?

DR. CARTER: So it sounds like the first question is for Dr. Udelson and the second question to Dr. LaVange.

DR. HARRINGTON: Just as he's getting up there, I know Dr. Tatum wants to get on this question. And then I'm going to go to Neil, who's been waiting for a while. So go ahead, Jim, and answer, and then we'll go to Dr. Tatum.

DR. UDELSON: Well, we don't have an analysis of a gold -- or an independent standard among all patients who had a disagreement of 2. The data I showed you just now about the angiographic standard if there was a disagreement was a disagreement between normal and abnormal. So we don't have anything at the moment along the scale anywhere about people greater than 2 apart.

DR. McGUIRE: You know, on CC-80, the histogram of the SDS deltas, if you could collapse the 3 and greater, everything from 3 to the right put into a single histogram or just present the numbers, that would be the same data that I'm interested in seeing.

DR. KOCH: Yes. That analysis hasn't been done. But the context of the 3 was that 3 was identified as the smallest magnitude that would be clinically relevant.

Larger values than 3 probably have greater clinical relevance than the 3 has. But the margin, when you're trying to define a margin like one and a half, you use what would be thought of as the very smallest magnitude of clinical relevance, so your margin is half of that.

Clinical relevance in terms of decision-making could be at 5. It could be at 8. Certainly I agree that if it's at 10 or more, that's more extreme. And the sponsor can do analyses based on the information that you see in this table that would look at criteria like 3 or more, 5 or more, 7 or more. And then it would become another judgment as to what should be the upper bound on that percent. Should it be 10 percent like it was for the extreme disagreements? Or would it be potentially more 15 percent, recognizing that for an upper limit to be below something, the point estimate has to be even smaller?

So for the extremes that were illustrated, the point estimates were down around 3 or 4 percent to assure that the upper limits were under 10 percent. But certainly the sponsor can do those analyses and share them with the FDA, as well as the FDA can do them themselves. And the same issue would apply to adenosine versus itself.

DR. HARRINGTON: All right. Let's go to Dr. Tatum.

DR. TATUM: I know that you said that on the hemodynamic data that you provided, the two groups were pretty much the same. But among the discordance between the two tests was the hemodynamic data any different?

DR. UDELSON: Let me just rephrase to make sure I understand. Among the patients who had some degree of discordance between binodenoson and adenosine, were there --

DR. TATUM: Or adenosine and adenosine, from two trials.

DR. UDELSON: -- were there differences in the hemodynamic blood pressure and heart rate response?

DR. TATUM: Yes.

DR. UDELSON: I don't know that, and I don't think we have that at the moment for you. I mean, theoretically we could put that together.

DR. TATUM: Because we know with vasodilators, reduction really could change things significantly.

DR. UDELSON: Yes.

DR. HARRINGTON: Neil?

DR. WEISSMAN: Thanks. I want to approach this categorical agreement from the image interpretation point of view. I, too, am very sympathetic about when a reference standard doesn't meet our expectations. In my world of echocardiography, we'll look to test/retest variability of echo to look at valvular regurgitation, which is done hundreds of thousands of times a day.

But when you look at that, about 25 percent of the time there's not exact agreement. However, what the disagreement is, is typically a one-categorical difference. So I sort of agree with Dr. Udelson in terms of not just classifying everything as normal or abnormal, but looking at those one categorical changes. And to me, those are reasonable expected variability, mild versus moderate, none versus mild.

What I had trouble understanding, and I would appreciate some help with this, is those 2 and 3 categorical changes because to me, a reading of none versus severe or severe versus none is a fundamental error that occurred someplace. And when that happens, you often go back to the primary data and do kind of a root cause analysis. And I was wondering if any additional analysis and insights into where this real gross variability has come from.

DR. HARRINGTON: So you're not so troubled, Neil, going from mild to moderate or maybe even moderate to severe. You're looking for something that really might change your opinion of the test moving a couple of categories.

DR. WEISSMAN: Correct. And I'm looking to see if we could have a better understanding of the issues, technical issues, variability issues, from those grossly --

DR. HARRINGTON: To see if you can explain within the raw data why that marked divergence might have happened?

DR. WEISSMAN: Exactly.

DR. UDELSON: That's a very important point. So let me reiterate that the images were read in a very blinded core lab, no clinical data. And there's an inherent variability, no matter how expert the readers, no matter how extensive the training. And these were expert readers with extensive training.

Now, after we had the 301 results, we did exactly what you suggested, and this is not very scientific, but it's an investigation. We took the most discordant pairs, and these guys flew up to Boston, and we sat and we looked at them, and with me. And I said, you know what? They aren't really that different. You know, I can see where may be the few 2s came over here, and maybe not. But they weren't that different.

The environment of looking at one image by itself and scoring segments with no clinical data is not really the environment of clinical practice, as you know. Moreover, we then after that subjective, biased, exercise, we brought in readers to look at side-by-side readings of the 301 data, similar to what we had done in 206, again, not in keeping with how you must analyze these by FDA guidance in pivotal trials, but to really understand was there a real difference?

So this post-hoc, side-by-side analysis of the 301 images produced kappa statistics between .78 and .92 for various readers. Now, you know, we cannot show you. I'm glad you asked the question, of course. But that is not in keeping with the guidance, the rigorous analysis. But on the other hand, it's a little more like clinical practice.

If you had a patient who had a SPECT study last year and they had one today, I'm looking at them side by side, and has this changed clinically? And it's somewhat subjective. I can use the 4DM SPECT program, et cetera. But the side-by-side reads in 301 as part of our post hoc investigation, the root cause analysis, as it were, as you suggested, showed much higher agreement than we had seen in the reads as done per guidance.

DR. HARRINGTON: Does that make you feel better or worse, Neil?

DR. WEISSMAN: Not a lot better yet.

So what you're implying is that it's mostly due to a reader variability more than a technical variability. But when you look at the inter-reading variability, it doesn't account for everything.

DR. UDELSON: It's a component. The inter-reader variability is a component of the issue with kappa.

DR. WEISSMAN: It gets to the second point, and it's for anybody. But we keep talking about the adenosine test as a whole. And the variability really is from the SPECT, probably moreso than just the stress agent. And we do on every single one of these patients have test/retest assessment of just the SPECT. It's called the rest study.

DR. HARRINGTON: That was Jim Neaton's point this morning. Right.

DR. WEISSMAN: Yeah. That's right. So, I mean, is there a way -- you know, I'll maybe address this to you, Mark Levenson, is, is there a way to look at that variability and subtract it out to be able to isolate the variability of the stress?

DR. LEVENSON: Well, in some sense, taking the difference is trying to do that, is trying to subtract off the most recent image to remove that variability. So when you take the difference between the rest and the stress, you're trying to accomplish that to some extent. It's the same imaging session. Other than that, I don't know any way.

DR. UDELSON: Well, maybe I can, Neil, show you the data.

So I'll put this slide up first. This is something just about your previous point, about the extreme differences, the last bullet here. And I'd mentioned that there were 22 patients with extreme differences in the upper right corner, in other words, severe adenosine abnormality, score greater than 8, and no ischemia on the binodenoson.

The first point is that only 8 of those 22 went on to angiography, which sort of implies that the site read suggested a much lower amount of ischemia. So again, core lab versus site reads, and there's literature on this in the nuclear world, that core lab reads show more ischemia sometimes than site reads, and the angiography tends to go the other way. So that's one point. And also, that of those 8, one agent was right half the time and the other agent was right the other half.

Can I have this slide up now?

Now, your other point was important in that you are correct. That was an internal -- some of the panel members this morning mentioned, and I think Dr. Levenson may have mentioned, or Dr. Marzella, that it was only in 305 that within the context of the trial, there was an adenosine/adenosine test/retest component for context. But all of the trials had rest and rest that you could pull out, as you suggested.

Then if we forget about the summed difference for a second and just look at this as an internal reference from all the studies -- this is 301 in the star, 302 in the gold, 305 in the green circle, and this is the sort of equivalence analysis with the margins of equivalence.

So here's the rest data. Here's the summed stress score, relative to that, so there's a lot of overlap here. And your point is, I think, that you can use the rest scores within the context of this reading environment to define the sort of inherent variability of SPECT imaging without any pharmacologic stress, without any B or A test agent. And so the summed stress scores -- and if you look down here at the computer analysis, sort of no human eyeballs, they are very overlapping.

If you can go to what I think is the next slide, the kappas -- thanks -- so again, if we ignore the difference for a second, so here is sort of the inherent variability with the rest/rest, versus a rest image done within a week, and the summed stress scores, a lot of overlap in the confidence intervals. And then down here, again remove the human element.

So, in essence, your question gets to a very important point of stripping away various elements of the variability. So strip away the element of the pharmacologic stress agent and the human. Strip away the element of the human here, and you see that they line up pretty carefully.

So that's an important point. There are many sources of variability. You know, we tried to remove in the protocol from the acquisition, the technical aspects of the imagery construction, and a lot of effort was made to remove those sources. They're hard to really measure. But some of these things, the agent and the SPECT image itself, can be looked at in that context.

DR. HARRINGTON: Are you okay, Neil?

DR. WEISSMAN: Yes.

DR. HARRINGTON: Let's go to John Flack. You were listed earlier.

Do you still have a question? Okay. And then you, Dr. Krantz.

DR. FLACK: I'm almost seasick listening, trying to figure this out.

DR. HARRINGTON: Drug approvals don't look so bad any more, do they?

[Laughter.]

DR. FLACK: The one thing, though, that I'm having difficulty with is trying to figure out how the people who got the cath with unblinded adenosine data, how they would be biased in favor of the new drug as opposed to adenosine.

Now, if you think about it, you're using that as the -- yes, of course there's bias. But why would that bias favor one drug over the other? If anything, the bias would probably favor the drug that is actually the data it's being used on, or you can argue that the bias that is reflected there is picking up the fact that this drug may actually be causing you to over-call the actual anatomic lesions, which are probably to some degree correlated with ischemia.

I would also submit, too, that just because these drugs cause the same amount of hyperemia, I think we're kidding ourselves if we really think we fundamentally understand that they could not actually cause differences in the amount of ischemia provoked in a ventricle. There may be things we don't understand.

Again, I'm not an expert in this area. But you never learn anything new if you know everything all at the same time. And I'm a little bit leery of just saying that they are actually absolutely the same even though they cause a similar amount of hyperemia. And in fact, there's probably enough variability in the hyperemia that it's possible you may get some differences there.

But specifically, is there any real reason to believe that the bias would be in favor of the new compound as opposed to adenosine? Because I can't figure that out. And when you actually get to the actual -- as imperfect as it is, and I'm not going to argue that point any more, it is the gold standard we're using. Maybe it's the copper standard. But at the same time, it's what we're using. And the gold standard, adenosine, did not do as well at predicting that as the new drug.

Also, a second question, is there any difference in the SDS scores in the cath group by whether adenosine or the new drug was proved correct in the discordant comparisons between the image and the actual cardiac cath data?

DR. HARRINGTON: Dr. Levenson, were you going to comment on his first point?

DR. LEVENSON: Yeah. I would like to comment on the potential bias in the angiography accuracy results.

Is there any way I can get my slide 24 up?

DR. HARRINGTON: Which one?

DR. LEVENSON: Twenty-four.

Okay. I'll try to do the best I can to explain where I think there might be bias here. I think it comes in two ways. If you look at the specificity of adenosine, it's very low. That's the 49.4 percent.

Since the judgment to go on to angiography is chiefly based on the adenosine, you're not seeing many negatives there. So if you don't see any negatives you're going to have a low specificity.

Now, the other place I see where there might be a bias is if the negatives do go on to angiography, there must have been some other clinical information that's making them go with that decision.

So the angiography, it's like if you see a negative in the adenosine result, there must be something very strong -- I mean, not as a clinician, I don't know quite what that would mean -- but there must be some other clinical information that's driving that patient on to angiography. So the negatives for adenosine that go on to angiography might not be a fair representation of the overall negative population of adenosine.

DR. FLACK: I don't discount at all that there's bias in how people got there. There's a tradeoff between sensitivity and specificity. But inherently, that's a characteristic of the modality you're using. And so it kind of is what it is. You got higher sensitivity and you get lower specificity.

But whatever it is you're using to send positives and negatives to the cath lab, what you're comparing it to actually beat it on the gold standard. And I still haven't heard an explanation that convinces me, past the bias in the sample, that the bias would actually favor the new drug in that and all. So there's bias there, but I'm not convinced that it favors the new drug.

DR. LEVENSON: Well, I would say the negatives for adenosine are not a fair representation, not a random sample of the overall population of negatives. They're the ones that if there's some additional clinical information, that's probably driving them to get the further procedure. If you just took a random sample of patients that received a negative adenosine, I think you would get a different result.

DR. HARRINGTON: Yes. I think you know what I would say, John, that this would be an example of since it was chosen and not random, there's going to be things we just don't know about. And so the only way you would really know is if you took a population who was scheduled for angiography and then randomized them to receive the tests, or to receive sequential tests and then be able to relate that to the angiography.

DR. FLACK: See, I disagree with that. I disagree with that in the sense that -- well, technically you're right. And when bias is operative and generating a sample, I'm still beating my head against the wall to figure out why the bias would actually work against adenosine here unless it was just something inherent in the way adenosine actually is giving you information.

So adenosine performs -- it looks like it picks up -- if you do a big number of people, it picks up a few more of the positives, okay, but you've got more false positives in there. Okay? Because your specificity is lower. Okay?

One of the explanations, outside of all the other alternative explanations, is this real? And it's the closest thing to real we actually have in here. Everything we're looking at is fuzzy. Every piece of information that I've heard today, there's questions about it.

This is probably the hardest piece of information we've actually got. Okay? Even the extreme differences between these two drugs basically don't look any different than the adenosine to adenosine. Okay?

So, yes, there is bias. But for me to be convinced it's working in one way or the other is what I need to really -- I need to go beyond the fact that, yes, there's probably bias in the data. But it's really here we're making a relatively contrast, and the relative contrast is between these two drugs. And to me, this is about the hardest evidence that we actually have because all this other stuff looks really fuzzy.

DR. HARRINGTON: So we've got three people who want to weigh in on the bias question. We've got Sebastian, Sanjay, Henry, and Mike. So let's go one at a time.

DR. SCHNEEWEISS: Okay. So this is Sebastian. So a quick clinical scenario is you have a patient with adenosine values. The decision to cath or not to cath is based on clinical factors and adenosine value. Right?

We know that cardiologists are not averse to catheterization, so they would rush this person to the cath. And coming from the Brigham, pretty much everybody's cathed, anyways. But here comes the misclassification part. For many of these patients, the adenosine and the bino value are fairly comparable. But there are some patients where the bino value is much lower than the adenosine value. And those patients would be, according to bino, classified as noncath or cath negative. Right? And adenosine they will be rushed to the cath lab anyways. That is why the specificity will be low in the adenosine value. But for those patients, they would drive the specificity high for the bino patients. Right?

So it's the combination of the cardiologists operating on those results and the misclassification together.

DR. FLACK: Is it not true that the majority of the people who got tested, even when they're positive, were not cathed?

DR. HARRINGTON: That's correct.

DR. LEVENSON: Only 15 percent of your sample, right, got cathed.

DR. FLACK: Only 15 percent of the sample got cathed. Okay?

DR. HARRINGTON: We know that from the Cedars Sinai data that was alluded to earlier that even in the highest risk group of patients, only approximately half of them get cathed. So it's not that highly positive tests -- that's why I made the comment this morning. There is not a logical linear line between a positive test and the cath lab. There's a lot of variability that goes into that decision-making.

I know that troubles you.

DR. FLACK: In relative terms, despite the biases that are there, you basically -- these agents -- really, to me, the greatest comparator is probably adenosine to adenosine. And there's probably not enough of that in this data set. But what we do have, it just doesn't look like this new agent is any worse than adenosine/adenosine. And when you get down to the real hard endpoint of cardiac catheterization, despite all the fancy explanations I've heard, I still -- this country boy from Oklahoma, I really don't see how you systematically bias it against adenosine by going to -- it's almost like saying, we've got a test, and if we actually use the data on the test, of course we're not going to be as good as the comparator because there's problems with it, is almost what it kind of comes across like. And to me, that's actually trying to have it both ways.

DR. HARRINGTON: Okay. Fair enough. And your exact point is the one we're going to get to when we get to the questions.

Go ahead, Sanjay.

DR. KAUL: I really don't have anything to add. I think Sebastian already --

DR. KAUL: I know you're not a simple doctor, though, from Oklahoma.

DR. KAUL: Clarify that -- I mean, this is the classic conundrum with post-test-referral bias, which inflates sensitivity and deflates specificity. And so there are methods that have been described. Beggs-Greene was the first one to describe it in '83, and subsequently there have been some simplified modifications, one of them by George Diamond, the other one from Ray Gibbons. And if it is possible to apply those tools to de-bias the data, then I would suggest to do it. But if only 16 percent of the subset underwent arteriography, is it really worthwhile doing that?

DR. HARRINGTON: Lyle, is your comment on the bias question?

DR. BROMELING: What percent of those who tested positive had angiography?

DR. HARRINGTON: Did you hear the question,

Dr. Udelson? If you were to do the binary positive/
negative, how many of the positives got cathed?

Is that your question? How many of your
negatives got cathed?

DR. KAUL: Right.

DR. UDELSON: I'm sure we have that. Hang on
a moment.

DR. HARRINGTON: While you're looking for
that, were there other comments on the bias issue?

Go ahead, Mike.

DR. DOMANSKI: Maybe it's not specifically on
the -- and the bias issue strikes me as pretty
straightforward. There's no question that it's
gross -- you know, it's data that you can either decide
the basis of the bias or its quantity from the data
available. I mean, that seems clear.

But I wanted to make a comment about Neil's
comment, and then I'll save other comments for --

DR. HARRINGTON: Could you hold that?

DR. DOMANSKI: Sure.

DR. HARRINGTON: Because I want to solve at
least this discussion.

Jim, did you have a comment on the --

DR. NEATON: Well, I just was going to say I think the bias question has been addressed accurately. I mean, I think if you had that slide up and you had adenosine twice, and you basically made a decision to go to the cath lab based on the second adenosine measurement, ignoring the first, you'd see the same result.

Essentially, what you're seeing is regression toward the mean. You're choosing out selectively higher on one of the measurements, more high scores along with all the other clinical evidence, and that's going to differentially affect sensitivity and specificity for another measure that was done simultaneously.

DR. HARRINGTON: Do you have the data, Jim?

DR. UDELSON: Let me just say not at our fingertips, Dr. Bromeling.

DR. HARRINGTON: Then we're going to keep going while you're looking for it.

Same topic, Darren? Okay.

DR. McGUIRE: Just a point of opinion is that even if we take away all the bias, what we're dealing with on that slide are 200 patients in each arm who underwent cath, so a very small cumulative sample size. Even if there were no bias, we're gravitating toward the truth standard here. And, again, back to my comments, it's possible that this will require a truth standard for comparison.

DR. HARRINGTON: So just to keep order here, we're moving off of this issue around the selection to the cath lab.

Mori, you're up next. Then we're going to go to Peter and Henry.

DR. KRANTZ: Is it too much of a digression to talk about safety?

DR. HARRINGTON: No. We can go into that for a bit. Yes.

DR. KRANTZ: Are you sure? Well, I'm certainly very confused also about the efficacy. I was looking at Study 302, and I realized that people I wouldn't cath would be those that were nonischemic or mild ischemia. And in both groups, it was 315 patients. I don't know whether to be reassured or frightened by that fact.

So the question I had about safety was, we heard a lot about symptom scores and whatnot. But what about needing to use aminophylline? Do we have any data on that?

DR. HARRINGTON: Good question.

DR. CARTER: Yes, we do. So in the three Phase 3 studies, there were a total of 6 patients that required aminophylline reversal, obviously on clinical grounds. And there were 2 on the binodenoson side and 4 on the adenosine side. And there were 4 issues such as chest pain, dyspnea, wheezing, and hypertension. So 4 adenosine, two bino.

DR. HARRINGTON: Jim, how does that compare in standard practice? Is that pretty typical? Infrequent?

DR. UDELSON: Well, with adenosine testing, when people know they're doing it, they actually rarely give aminophylline because they know that when they turn the infusion off, whatever's happening will be over, very occasionally. So this was double-blind, double-dummy, so the clinicians were reacting to symptoms they were having. And then the issue with binodenoson, of course, this defines those data. So it's 2 out of 1100 or so.

DR. KRANTZ: It just seems awfully low to me. And I think clinically we use more aminophylline. And maybe we're using it too much. But I do wonder, though, symptom scores that are relatively subjective. And if people are really sick, you'd think there would be a much greater amount of adenosine.

DR. UDELSON: To use adenosine?

DR. KRANTZ: Aminophylline. I'm sorry.

DR. UDELSON: To use adenosine testing for pharmacologic stress? In other words, you're saying you use aminophylline to reverse the side effects with adenosine? Or do you use dipyridamole?

DR. KRANTZ: Yes. I agree. We use the aminophylline much more with dipyridamole. I just wonder if we could have a more objective way of assessing symptomatology.

DR. UDELSON: Right. I mean, the -- no, now I see what you're saying. I mean, the idea was to capture rigorously the side effects. Now, no matter what the patient was getting -- the binodenoson was given as a bolus and adenosine was an infusion -- of course, the investigators knew that one of them might have been adenosine. And thus they knew that as soon as the six minutes were over, you know, a minute later all the effects would be gone. And so they obviously didn't reach for aminophylline, only 4 percent of the time.

But the symptoms were captured in a double-blind, double-dummy study, prospectively defined, training of the investigators, validated tools, because the purpose of developing this type of agent is to reduce side effects.

DR. HARRINGTON: Go ahead to Peter, then Henry.

DR. CONTI: I was originally going to ask if they had done the side-by-side comparison with the other trials. But I think they did say that they've done the 302.

Is that correct, just 302? Oh, 301?

It might be informative to do 305 as well because that again gives you adenosine versus adenosine. And it might be helpful not only from the perspective of having exactly the same tests done again and doing the side-by-side comparison, but also from a training perspective, if you were to do these in a blinded fashion, you could do adenosine/adenosine and then go on to do the two-arm trial components of 305 from there and see how that compares to 206 and 302.

DR. HARRINGTON: Henry?

DR. BLACK: I just want to make a few general comments. I think this data has been tortured beyond description. It's been waterboarded, at least.

[Laughter.]

DR. BLACK: And I don't think we're going to get anything more out of it than we've -- it's worse than fuzzy, John. I think it's uninterpretable as far as efficacy goes. It seems to me you tried to combine an efficacy study, as how good this was, with an effectiveness study, about what people did with the information, and I don't think that's a really good way to go ahead with it.

I think the one thing I'm reasonably sure about is that it seems to have a better side effect profile than something we use. I don't think I'm convinced that it's better and I don't think I'm convinced that it's worse. But I am pretty much convinced that it's a safer agent and a better tolerated agent.

What I remember about how a screening test ought to come out is it ought to have -- you ought to err on the side of being more sensitive and sacrificing specificity. We don't have the ideal way to do it, but that's what we got. And we can't probably apply the kinds of things you were talking about because the data has been collected on a very small number of people. So I don't think we can improve on that, either.

So I don't know that we're going to find much else out with what we got except that it seems to be a better tolerated agent.

DR. HARRINGTON: So I've got Sebastian, Mike, and then Sanjay. And then if no one else has questions, we're going to break and then come back and go through the specific questions. So ask your questions now, in the next 15 minutes or so.

Go ahead, Sebastian.

DR. SCHNEEWEISS: All right. In light of what Henry just said, this is almost moot. But nevertheless, I want to emphasize that the sponsor had shown us data of the use of these tests in routine care. And the, by far, largest population are those patients with known CAD, which is most reflective of Study 305. So we'll get back to Study 305. So if nobody wants to hear this any more, tell me and I shut up.

When we go to slide CC-72, what these data try to tell me is, if you look at Study 305, the point estimate of bino versus adenosine is statistically significant, different from zero, and the measure being the mean SDS, where most of the people in the room here agree that it's a centralizing metric. Right? It nevertheless becomes statistically significant.

Don't get me wrong. I'm not writing on P values here. But this data is trying to tell me something, particularly in light that the point of adenosine versus adenosine is almost zero. Right? And if you look at the computerized data on CC-78, the difference is extreme, more extreme for Study 305. So I was wondering what it is, how the sponsor is trying to explain this.

The other point that I have is the bino

versus adenosine in 305 is not reaching the specified cut point of 1.5. The study number 301 showed a mean standard SDS difference of 0.15. That's ten times smaller than this cut point. The cut point, the chronology is what's defined after the results of 301 were available. A cut point of ten times larger than what was found in 301 was chosen.

Now, don't get me wrong. I believe the sincerity of the sponsor's page number 51 document, how they came up with the 1.5. But this should provide a peaceful sleep for the next coupe of years, I would think, because a ten times higher threshold of what you observe already at this point when you define the threshold is hard to beat.

So it's kind of two questions here.

DR. CARTER: Can I just sort of comment on that last piece? Again, let me just stress that we didn't back into these clinically equivalence margins. We carefully justified them on the basis of clinical relevance and the fact that we see what we see based on the data.

Actually, none of us have been sleeping very well at all for about 12 years, those of us who have been on this project that long. So no, this was very much prospectively defined equivalence margins that were done with a clinical rationale that we've tried to explain in some detail.

DR. HARRINGTON: Dr. Koch?

DR. KOCH: Yes. The margin governs where the confidence limits fall. So the margin isn't really related to the point estimate. The point estimate does need to be close to zero with a small estimate of variability for the confidence interval to be successful. But you typically wouldn't have a margin comparable to a point estimate near zero because that wouldn't account for what the variability would be. So the margins were based on both needing the point estimate to be near zero, which is the case here -- most of the point estimates are near zero -- and to have the length of the confidence interval sufficiently narrow that it would be entirely contained within the two dotted lines.

Now, you did note from slide 72, if we go back to 72, that the confidence interval from 305 is slightly to the left of zero, which would correspond to a significant difference. But still that confidence interval does lie entirely within the pre-specified range of minus one and half to plus one and a half.

We did address that concern with slide 81, which we've talked about previously, where we basically focus on at the mean of the absolute values and compare the bino-adeno difference against adenosine versus itself. And there again, we do get a confidence interval on means of absolute differences that is within minus 1 to plus 1. And that's the sense in which we found some reassurance relative to the tendency in slide 72 for the Study 305 to have an interval that was slightly to the right, as you had noted before.

DR. HARRINGTON: Okay. Let's go to Dr. Domanski, then Dr. Kaul.

DR. DOMANSKI: I just want to -- because it may be part of the discussion as we answer the questions.

Neil, I want to say something about the MR analogy. What you say about mitral regurgitation is true, but I would argue that the decisions that are made that relate to mitral regurgitation and the whole treatment paradigm is different than ischemia. So I'm concerned about arguing by analogy and would suggest in this matter we do it from first principles.

DR. HARRINGTON: Fair enough.

Sanjay?

DR. KAUL: The question I have for the sponsor, and I'm trying to sort of address the unmet need issue, what advantage does bino have over rega?

DR. HARRINGTON: I'm sorry. I didn't hear you, Sanjay.

DR. KAUL: Over regadenoson?

DR. HARRINGTON: Oh, okay. Well, of course they didn't do that study.

DR. KAUL: I know. But I'm trying to sort of get my head around unmet need. We agree that the tolerability is improved, and I have lingering questions regarding whether it provides equivalent diagnostic information. So the question I'm trying to wrap my head is, is there really an unmet need? And if there is, how does it offer that?

DR. HARRINGTON: Remember, in fairness to the sponsor is that there's no regulatory hurdle, for example, that they have to go against another agent. And they've been on this development path, it sounds like, a long time. But you're still wondering what does this add?

DR. KAUL: Exactly.

DR. CARTER: From the sponsor's perspective, although I may very well have a point of view here, it's not appropriate at all for me to comment at all on the performance and the qualities and everything else of regadenoson. So I really cannot give you an opinion here in terms of that.

Our intention was to come up with an equivalent diagnostic tool, if you will, with a better tolerated and a better safety profile. That we believe to be the unmet need. And we obviously believe that we've met that.

Jim?

DR. UDELSON: So, Dr. Kaul, my understanding of the primary endpoint of the regadenoson analysis was a noninferiority analysis of the exact agreement along the diagonal. So it was about 63 percent only for the adenosine/adenosine, and regadenoson was not inferior to that 63 percent.

If you look, however, at least in the published data, which is what I've seen, the difference, I think, is in the side effect profile. Dyspnea was numerically higher with regadenoson than with adenosine. Some of the others were lower. A composite score was slightly lower, statistically significant, but the components went in different directions, which undermines the strength of the composite.

So if you take the published data -- obviously the FDA has seen much more than that -- and compare it to here, I think it would be fair to say that the side effect data are -- the tolerability data are stronger here than they are for regadenoson. And perhaps that's reflected in the regadenoson label, although I'm not sure about that.

DR. HARRINGTON: Go ahead, Darren.

DR. McGUIRE: Just carrying along those, trying to envision the clinical niche, the bronchospasm, the asthmatic study that was uncontrolled, are those the only data? I think on CC slide 32, the fifth bullet suggested that there's decreased risk for bronchospasm. But in the absence of comparators, is that decrease compared with historical expectation or should that be worded as low potential?

DR. UDELSON: Well, my understanding of the 203 study was that it was controlled. And I think there was a saline control.

Rich, is that correct? Yes? So it was a double-blind saline control. So it was a controlled study, with just binodenoson showing no change in pulmonary function tests.

Now, this was, as it says there, in mild asthma. Again, I won't speak for the sponsor, but if I was an agency, I'd want a further degree of people with moderate asthma studied before I entertained the possibility of putting that in a label.

So I think this was a first step toward the possibility of using it in that important group of patients for whom adenosine is contraindicated. And what we do in practice, as I'm sure you know, is we go on to dobutamine, which is very difficult for patients; it's difficult for clinicians. So at the moment it's a potential, given the selectivity.

DR. McGUIRE: Okay. And just for semantics, so this decrease should really say low potential or no observed potential. Decrease suggests that it's better than some comparator, and I'm certain it wasn't better than placebo.

DR. UDELSON: That would be fair.

DR. McGUIRE: Okay. I just wanted to be sure we weren't missing some other comparative data, even from the randomized trials.

DR. HARRINGTON: All right.

Dr. Udelson, before you sit down, I'm going to ask the last question before the break.

I'm always intrigued when I read the different briefing books when there are statements that say that the FDA suggested one thing and the sponsor did another because usually the sponsor marches in tune with what the agency asks.

The design that they had suggested, and correct me if I'm reading this incorrectly, would have been take a group, and you were going to test, A versus B, and then for the retest it would have been again randomized A versus B, so that you could do all of the various inter-agent variability as well intra-agent variability, but the sponsor elected not to follow that advice.

Was that on recommendation of the steering committee, that they did not feel that was an appropriate design? Help us understand that.

DR. UDELSON: Okay. Well, part of it has to do with the trajectory of the timeline. And maybe we could have that timeline slide up. When the 301 -- well, I'll give you my opinion and then the sponsor can give you theirs, from their point of view because that may be different.

Can I have the slide up, please?

So the initial pivotal trials or Phase 3 trials were designed prior to the publication here, the FDA guidance document, which it was the guidance document that suggested, take patients coming in who are having a test such as adenosine, do the adenosine test, and then randomize them to have either the new test or the adenosine test again. And that was the design of the regadenoson study.

So these studies were designed prior to that. Now, you might ask, well, why didn't we do that, follow that guidance, for the 305 study? And we certainly considered that. But that would lead to then problems combining the data, the comparability of the 305 study, particularly the adenosine/adenosine arm with the others, which we envisioned would be important.

Then there's one final point. The design, as suggested in the guidance, is robust when the major endpoint of interest is the imaging data itself without anything else. The additional dimension here is the side effects. And, again, I keep returning to the fact that the only reason to develop this class of drugs is to lower side effects. If it increases -- you know, I agree with Dr. Bengel. If it increases coronary blood flow to a similar degree of adenosine, the rest of the problems are just SPECT imaging problems. It's not a binodenoson/adenosine-regadenoson problem, but it's about the side effects.

It was our opinion that the regadenoson design is a parallel group design. So you are comparing side effects in different groups of patients; whereas here, because of the crossover, every patient is compared to themselves, which makes the side effect data even more robust.

DR. HARRINGTON: That's very helpful.

All right. I'm going to look around.

Any final questions for the sponsor or for the FDA? Because if not, why don't we break for about 10 or 12 minutes, and then we'll come back and go through the questions.

(Whereupon, a recess was taken from 2:57 p.m. to 3:14 p.m.)

DR. HARRINGTON: All right. If I could have people take their seat. We're contrasting these questions, Dr. Rieves, to what we'll see tomorrow with Dr. Stockbridge, which will be a crescendo approach to the questions. But in this one, we're actually going to vote on one question.

So there's four areas that Dr. Rieves and the division would like us to discuss. And on the last one, we'll vote, and Elaine will have me read the voting procedure before we officially vote.

So can we put the first question up there, Elaine, or is it just reading it?

MS. FERGUSON: No. I got it.

DR. HARRINGTON: So the first several are a discussion. And what we'll try to do is have a robust enough discussion that we either, as a group, move to some sort of consensus, or move to at least a series of points that the FDA can take away and understand where the advisory group at least stood on it. And that can certainly be a majority. But I think, importantly, in these discussions is to make sure that the minority opinion is also heard if we don't have consensus.

So the first discussion point is, the primary endpoints for Studies 302 and 305 were changed from a patient-level concordance of binodenoson and adenosine myocardial perfusion images, or MPI, to a comparison of average summed difference scores.

Do the revised endpoints provide a robust measure of agreement between binodenoson and adenosine MPI?

So I'll open it up to whoever would like to start. All right. You know I'm just going to pick on somebody.

Go ahead, Henry.

DR. TATUM: I'd like a definition of robust.

[Laughter.]

DR. HARRINGTON: Why did I know that was coming?

Dr. Rieves, would you like to provide a definition of robust?

DR. RIEVES: These initial questions really are meant to be somewhat provocative of the discussion. So the actual wording is probably not that critical. But in general, I think the question relates to, does it improve the assessment of agreement between the test and comparator compared to the original kappa statistic? Is it a better statistical comparison measure?

DR. HARRINGTON: Well, could I interpret it maybe even another way, is that you obviously gave guidance and they launched their first study, and you were content with their measurement of the kappa statistics as a way of looking at the referenced comparison.

Is that a fair statement?

DR. RIEVES: Looking back over it, we did not object to that, right, because it was conceivable. It was conceivable that study design could have been very successful. The success is data-driven, if you will, by the results. It was conceivable. We did not object to it.

DR. HARRINGTON: So is one way of interpreting robust, that if that was something that you did not object to, that the new proposal should be at least as unobjectionable as that or better?

DR. RIEVES: That's true. Hopefully better. Hopefully better.

DR. HALPERIN: As I've heard the lengthy and very, I think, comprehensive discussions of the various methods of assessing concordance, I think that Dr. Unger made, and then Dr. Harrington ratified, a very important point.

We're essentially looking here at an active control comparison of two diagnostic agents. And whether you regard it as a noninferiority or equivalence comparison, essentially what we need here is some external standard that establishes the quality of the assessment method the same way we would in an anticoagulation trial where we'd be comparing with the adequacy of standard therapy.

As you pointed out, if all the pharmaceuticals were either dropped on the floor or if the camera was shaking a good deal and none of the images could be discerned at all, we have clearcut comparison and equivalence.

So where can we look in these data for some quality measure? And the only place I can think to look is in the adenosine/adenosine comparison for the patients that had perfusion defects because this is a compound that's designed to reveal perfusion defects. And there I see 14 patients with perfusion defects in which there was concordance. And I'm bothered by that lack of power.

Just a comment, and I would be very eager to hear comments from others about that.

DR. HARRINGTON: So this gets to a point, I think, that Dr. McGuire brought up a bit ago, which was that the bulk of the data that informs the data set are from the normal patients. No perfusion abnormality.

DR. HALPERIN: Precisely.

DR. HARRINGTON: And you're saying if you remove that, we're actually left with very little data in which to draw our inference.

DR. HALPERIN: Particularly when it comes to the quality measure, which is the adenosine/adenosine comparison. That's what tells us about how good our assay is to evaluate differences or similarities in the treatments, or in the diagnostic compounds, rather.

DR. HARRINGTON: So if we took that, Jonathan, as a broad statement, and that the lack then of sufficient numbers is bothersome to you, that there's just not enough information, then force yourself to look at the specific question. Does it bother you, having changed -- you know, to prove or to determine noninferiority, does it bother you having changed from one methodology to the other, or does it more bother you that you just don't believe there's enough information at all in the data set?

DR. HALPERIN: I think what I'm saying is that -- and you get to the issue, which is although there are many ways to look at the boundaries of confidence here, the real question is can we trust the assay at all. And the answer to that question in this data set can come only from the limited number of abnormals in whom we have an adenosine/adenosine comparison because everything else is, frankly, statistical gobbledygook. We have boundaries that can be defined in many ways. But ultimately, all that we will differ or agree upon is to what extent we can feel comfortable that they have shown equivalence. And the question about equivalence comes down to the adequacy of quality in the assessment. And that draws itself to the issue of abnormal detection in the standard way.

DR. HARRINGTON: That's helpful.

Mike?

DR. DOMANSKI: I think the answer to the question is no. The problem is that in a given patient, having a normal study or not having a normal study decides, in effect, in practice, whether the patient goes to the cath lab or not. At least, that's the usual practice. And, in fact, if you look at it that way, this seems to be different from the adenosine about 20 percent of the time.

So I think the answeris no. And I think that the angiographic data presented to impugn the reference standard that the sponsor themselves put forward is, fortunately, probably, for the sponsor, not interpretable. So I think the short answer is no.

DR. HARRINGTON: So let me push you as well a bit. Granted that most of the patients within the data set are normal, the reality is, in most of nuclear cardiology practice, that's who's being studied. And so having a normal scan actually keeps you out of the cath lab, which is probably -- even as a cath lab doctor, that's probably a pretty good thing.

DR. DOMANSKI: Yeah. I think as a cath lab doctor, I'm more worried about missing disease than I am about whether I do a very low-risk catheterization procedure. And it looks like you miss it 20 percent of the time with this, if you accept adenosine as a reference standard. I mean, you can't both accept it as a reference standard in your pivotal study and then try to impugn it. If you impugn it successfully, you need a different reference standard.

DR. HARRINGTON: Okay.

Other comments around the table? Jim and then Dr. Bengel?

DR. NEATON: Well, I think the answer is, in part, you need both, in my mind. I mean, if I saw the Study 202, I think was the number, with all of the people in the non-ischemic cell, I would certainly stay away from the kappa statistic as the primary analysis of what I did because it's just going to be, you know, impossible to achieve the .61 that was kind of laid out there.

DR. HARRINGTON: So that's a really important statement.

DR. NEATON: So I just --

DR. HARRINGTON: You were okay with what the sponsor did in saying, now that we've got more data, their kappa statistic of the lower confidence interval hitting .61 just wasn't realistic, and you accept that?

DR. NEATON: I think the problem with both approaches that I've seen is that I don't have a good sense for what the bounds of noninferiority should be, so that the .61 was derived not based on any clinical basis at all, from what we heard this morning. It was basically derived because they observed .75 with one method of comparison in the early study, and they thought it would be reasonable then to hit the .61 in their pivotal studies.

I think there's advantages to both approaches. I mean, you're right about the -- if you throw all the scans on the floor and you pick them up and you get zero, that's not good. So you can't focus just on the difference. You have to focus on the standard deviation of the difference as well.

Now, so a major limitation that we're working with is they only have that data in the one study. And so we have data on adenosine/adenosine concordance from one study, not all three. But to the extent that we have, those standard deviations are similar to one another. And so that I would probably kind of want to look at both, overall agreement, once I understand kind of what the bounds of agreement would be, but also focused on the continuous range of the score that they looked at.

It seems to me that I'm still not real clear on kind of what's being judged here as far as kind of what's ideal, and so that Mike argued this morning that what you don't want to do is send somebody to the cath lab unnecessarily or miss somebody that's really got an important defect.

So while you cannot put these scans side by side and make that judgment because that just would be an inappropriate design, I don't see why you couldn't put them side by side if you mixed up a whole bunch of other scans with them that were, you know, the same patient from other places or even different patients, and have a judgment made, is what I'm hearing, a judgment made based on the clinical data plus the scanned results as to whether they should go to the cath lab or not. And we don't have that information as to whether that agreement -- there's disagreement on that point. And I asked for it but I haven't seen it, the data to justify the 1.5 based on the data from the prognostic studies.

I mean, there's an error, I think, in the report at what's cited there. And so surely that could be generated. And so that my estimate is that the difference that they cited is associated with a 15 percent increase in cardiac disease. I think that's pretty high in terms of for tests like this. And so I actually would have probably set the bound smaller than 1.5.

DR. HARRINGTON: I'm going to go to Henry next, but I'm going to push you a little bit to help us on the statistics here.

One of the fundamental issues here is that the FDA statistical group does not agree that the average summed difference score is a robust enough measure of trying to compare the two things. And I think what I hear you saying is that the approach that the sponsor took -- I think they call it the totality of the data approach -- that they're showing us that. But they're also showing us the standard deviations. They're showing us a lot of different pieces of information which they are saying are supportive.

DR. NEATON: I think that's perfectly appropriate. I do. And what I'm lacking, and what --

DR. HARRINGTON: So you're not bothered by the average summed difference score as long as it comes with some other things?

DR. NEATON: Right.

DR. HARRINGTON: Are you bothered by it as a primary endpoint?

DR. NEATON: It's not an appropriate primary endpoint with looking at the standard deviation along with it or by looking at percent of major discordance by some criteria like they propose. You can't look at it by itself.

DR. HARRINGTON: So you need it --

DR. NEATON: You need it in connection with --

DR. HARRINGTON: -- with other things.

DR. NEATON: -- another parameter that should be looked at.

DR. HARRINGTON: And again, from your perspective as a statistician, Jonathan's comment that in the group of -- if you eliminate the normal, so to speak, the sample size is pretty small. Does that --

DR. NEATON: Yeah. I think it is. And I'm kind of torn between, you know, that observation, of course, and the fact that this is just the way it is. This is the way real life is in terms of the kind of people that come in and get this test. And making errors on those normals is important. And so, another way of turning that around is you want to have a fair number of people there because maybe a very bad error is to send a perfectly normal person to the cath lab.

DR. HARRINGTON: Henry?

DR. BLACK: Yeah. I want to follow up a little bit on that and, you know, play internist and referring doctor here.

I'm sending a person for this test because I'm not sure whether they have coronary disease or they need revascularization. If I'm sure they do, they're going to go right to the cath lab. My prior probability is going to outstrip any sensitivity you could possibly get.

So I think they studied the right population for what I think a screening test ought to be used for. So I'm not sure we can eliminate the normals or the low-risk people, necessarily. The intermediate risk people I think are the most important group, and they did weigh the sample, so there were a lot of them.

I again go back to look what the options would be. Some people would just cath anybody you had an intermediate suspicion of. Others would do something else to try to avoid the cath. And you're still not going to really get the answer, I think, until you do that. And it doesn't even tell us what we really want to know, which is what the angiographic findings mean with respect to outcome. We're far away from what we're really after when we screen people.

DR. HARRINGTON: So let's go specifically to the question at hand, Henry. You know, they had one endpoint. We heard a lot about the comparison between the kappa statistic and other methodologies of testing the comparability between the two tests.

Do you think this is an arcane argument and it's not helpful to you or do you think that you would put your vote down on one or the other as one of the tests being a preferred choice?

DR. BLACK: Well, I'm not bothered by someone, when data was still blinded, deciding that they had made a mistake. I know what I would do if I were there. I'd ask Jim Neaton what he would do and leave it at that.

DR. HARRINGTON: And I think that there was a lot of discussion throughout the day, and most of the people around the table, I think, have said, well, you know, the more data became available, they changed their analytical approach.

DR. BLACK: I mean, they're to be congratulated for how they trained people, how they read the studies, what they did when they saw what they were planning, and all this energy about might not giving them an answer. There's a lot of things that were good. But, still, now that that's done, I don't think they should be held accountable for not making a midcourse correction that seemed necessary without ruining what they had planned.

DR. HARRINGTON: Okay. Fair enough.

Sanjay, and then Lyle.

DR. KAUL: The answer is no, and let me try to justify that. First, the sponsor chose the least burdensome pathway, which is, from a regulatory perspective, still acceptable. Their justification for changing the endpoint did not persuade me. And when you compound that with the lack of an internal control in two out of the three studies, and more importantly, lack of adequate number of abnormal scans, I have no way of predicting what impact that would have in either limiting or inflating the moderate or extreme degree of discordance.

So for those four reasons, the answer to the question is no.

DR. HARRINGTON: So let me also push you, Sanjay. You said they chose the least burdensome path. That's not what they described that they did. They described -- you know, I think Dr. Carter said we didn't back into this. We looked at the current field of evidence. We made some assumptions based on data that were available in the field about what's an important ischemic size, and we built our analysis around that.

Yes, it was now overpowered relative to what the kappa statistics did to the power calculations. But why do you call that least burdensome?

DR. KAUL: Let me clarify. There are two ways the FDA will allow them, in terms of the efficacy. One is a comparison to a truth standard and the other one is a degree of agreement. And it is my opinion, and I think Dr. Rieves pointed that out, that the more optimal way of coming to an efficacy assessment would have been comparison to a truth standard. That's what I mean least burdensome.

DR. HARRINGTON: Is that a true statement, Dr. Rieves?

DR. RIEVES: I think that's basically true. One alternative would be, of course, to have clinical outcomes as a truth standard, and develop it as a diagnostic with prognostic ability. That is always on the table. Your point's well taken.

DR. HARRINGTON: I'm not done with you yet, Sanjay.

You said that they did not provide sufficient justification or persuasive justification to change their endpoint. Henry says that they learned stuff along the way and, you know, give them credit. TThey took that into consideration and redesigned their analysis plan in a proper way. They were still blinded. They didn't have knowledge of the treatment comparisons.

Why do you not find that compelling?

DR. KAUL: Well, as I said in my first comment that I made, that I'm sympathetic to their predicament. They overestimated their kappa statistic because of many reasons, some known and some others not known. And so, as happens in clinical trials, you change your endpoints sometimes, but you are able to justify it. And the justification that I heard from both statistical as well as clinical perspective did not persuade me.

There's a lot of uncertainty in picking a 5 percent perfusion abnormality in isolation, not keeping the clinical context in mind. If I had seen how the data would have panned out in the subset of diabetics or the subset who had systolic dysfunction, perhaps I could have been persuaded a little bit more.

I also remain not persuaded with regard to the statistical reasoning. I'm not quite sure whether it applies here; I haven't really looked at it carefully, but if you have a metric that has a wide variance and you take 50 percent of that variance as your limits of equivalence, I think it's arguably not a robust or a conservative estimate. I would have taken 25 percent. And I did not hear a persuasive argument why 15 was more preferable than 25 percent.

DR. HARRINGTON: Again, just a little push back. Several people brought up that this sounds to be a safer agent, potentially.

Does that weigh into your mind when you start thinking about what's a persuasive level of uncertainty? In other words, would you give up 25, 50 percent if it's safer? And does the discussion of safety that we had, bronchospasm, AV block, does that matter to you?

DR. KAUL: Yes. It does matter quite a bit. What is the maximum loss in efficacy that is acceptable given the ancillary advantages? And I think we tried to encourage the sponsor to provide a concrete statement with that respect, and I did not hear that.

DR. HARRINGTON: Okay. Fair enough.

Lyle?

DR. BROMELING: I can see why they changed from kappa to the SDS. And it seems reasonable to me to use the SDS score. However, I'd be more confident in the use of the SDS score if they would have justified the equivalence constant, namely, plus or minus 1.5, by a formal statistical argument, where they would have stated a null hypothesis, an alternative hypothesis, and they would have given a power analysis or a power curve for interesting alternatives under the alternative hypothesis of equivalence.

Now, that was alluded to somewhat. I think they mentioned something about 95 percent power, but that was in conjunction with the kappa. Right? I didn't see anything in their document for a power curve justifying the sample size.

DR. HARRINGTON: Now, I thought -- correct me if -- maybe Jonathan Halperin remembers this because he asked the question about power. I interpreted their remarks that the original power calculation built around the kappa statistics as the test statistic was that they had 90-plus percent power at the 05 level.

When they switched their methodology, they now had power in excess of 95 percent.

DR. BROMELING: Yeah. But I haven't seen a power curve for differences under the alternative hypothesis.

DR. HARRINGTON: So to understand --

DR. BROMELING: It would be much more convincing to see a graph.

DR. HARRINGTON: So if they had shown us varying levels of where the boundary would be --

DR. BROMELING: Right.

DR. HARRINGTON: Okay. And then calculate --

DR. BROMELING: For various levels of equivalence, they can compute a power. I'd like to see those values.

DR. HARRINGTON: To give you a measure of surety as to how much you're willing to trade off?

DR. BROMELING: Right. If they were high enough, I would feel better about their choice of the SDS, the paired difference.

DR. HARRINGTON: So help me out with the question that Mark has brought up and that the sponsor tried to refute a couple of times, where Dr. Levenson brings up this notion that if in some patients you've got a minus 4 difference and other patients you get a plus 4, and then in other patients you get a 1 difference, when you mean all of that difference, you're only at 1, or less than 1, actually, .3 or something.

Does that bother you? It seemed to bother Dr. Levenson a lot.

DR. BROMELING: It does bother me. But I thought they also mentioned that they were considering the extreme cell in that 4x4 table. That wasn't mentioned, by the --

DR. HARRINGTON: And that helps to alleviate some of that risk. Okay.

Go ahead, Frank.

DR. BENGEL: I think I probably have to take a slight opposite position as compared to what has been discussed, just to also bring up the other side. I think this question is difficult to answer, do the revised endpoints provide a robust measure?

If we think about this in absolute terms, it's certainly debatable. But if we think about it in relative terms as compared to the initially defined primary end points, I would probably say that they are robust because the initially defined end points are also based on assumptions, and these assumptions are, in my eyes, at least, retrospectively seen not very realistic, either. They are based on a Phase 2 study, on a side-by-side comparison of images, where we have repeatedly said that this is inappropriate to do in a Phase 3 study.

So I would think that probably what has been learned after this 301 study was that the initial assumptions were way too stringent to come up with any kind of meaningful results for the upcoming Phase 3 studies. And this was the rationale behind adjusting the endpoint for the following Phase 3 trials, and I think the way this was done was a practical way, and I would think also, from a clinical perspective, was a reasonable way.

DR. HARRINGTON: So if I were to use the words that were on the screen here, are you saying that the first test, a kappa test as set out based on the Phase 2 data, which you say -- in which they admit -- as Jim Udelson said, we know that that's not how we're going to do it in Phase 3, but we were picking a dose and trying to understand some things. And so they probably overestimated the agreement.

So would you say it was too robust the first?

DR. BENGEL: I would think that probably at that time point, my interpretation would be at that time there was too much enthusiasm about the power of the technique. And since that time, it wasn't only the 301 Phase study. It was also other studies, including the regadenoson trials, which have shown that with this kind of an ambitious approach, you may not be able to obtain any meaningful results. That's why the endpoints were adjusted.

DR. HARRINGTON: So the second endpoint that was chosen, without putting words in your mouth, could I say that in your view it was a reasonable measure of agreement?

DR. BENGEL: Yes. Yes.

DR. HARRINGTON: Okay. Go ahead, Sebastian.

DR. SCHNEEWEISS: Here's a proposal for improved metric, which is you just take the mean of the absolute differences, which by that you lose the directionality, obviously, but you preserve the variation. I'm not sure whether that has published or -- I'm sure somebody has thought about this, so that would be easy to run in your data.

Otherwise, since you should provide advice for FDA, I think FDA should maybe more consider the vascular flow, as Dr. Bengel had mentioned already, because there's so much variation, because you're really testing or evaluating clinical strategy, which is the drug, which is the scan, which is interpreting the scan, which is summarizing the scan.

There are lots of sources for variation. So why not disentangle those sources of variation? First look at the drug effect itself because we understand the biology fairly well here, and look at the flow. So considering that in the larger picture and then summarizing all the evidence that is out there.

DR. HARRINGTON: Other thoughts around the table?

Go ahead, Emil, and then Mori.

DR. PAGANINI: Just a quick -- I'd do a little wordsmithing on the exact question.

Do the revised endpoints alone provide a robust measure? I'd say no. Do the revised endpoints, with other data presented, provide a robust measurement? I'd say yes.

DR. HARRINGTON: So you're in the Jim Neaton camp.

DR. PAGANINI: I think that they saw a

problem. They tried to solve it. The problem that they solved it with, especially the average, creates a lot of problems. They recognized that, and then they went and did other analyses to try to combat that.

I still have a real problem with MPIs. And so we'll get into that in other questions. But that being said, if we're just looking directly at this, I think that if it's alone, just the average sum, I think FDA is absolutely right; it's not adequate. But if you use that in the mosaic of everything else that they've presented in data analyses, I think then it becomes relevant.

DR. HARRINGTON: I think that's consistent with what Jim Neaton's saying, that one measure by itself is not robust, but a lot of additional analysis. And you, like several others, are not troubled by the mid-course correction.

So respond to Sanjay, who says, it just wasn't persuasive enough to change course midway through.

DR. KAUL: Let me clarify. The data from 301 was persuasive enough to change. The change that they made, I'm not persuaded by the justification of that change.

DR. HARRINGTON: Fair. Go ahead, Jim.

DR. NEATON: Actually, I agree with Sanjay on two of the points that he made. I mean, one is the -- I don't understand the 50 percent of the standard deviation. And I would rather see this based on a clinical basis, which we've asked for and haven't seen but should be able to get obtained.

But the other point that you made, which I think is very important -- and presumably they can do this, too -- is that -- correct me if I'm wrong, but if you have a low-risk person, based on all other clinical factors -- they're not diabetic and all the other -- they don't have dyspnea and other factors -- and you miss a defect, that's pretty bad, I would say, because that's somebody, maybe something that you might could do something about if it's an important defect.

So I just think the understanding, the disagreement, by underlying patient risk is important to do.

DR. HARRINGTON: Well, I think this gets into

Neil Weissman's point earlier, which is how much of a defect do you miss? If you're moving slightly along the scale, maybe it's not so bad. It's when you really miss it, you know, severe versus none.

Go ahead, Mike, and then we're going to go to Mori.

DR. DOMANSKI: Yeah. I want to underscore the fact that missing it by a little means a big difference in the clinical course of the patient. I mean, either you go to the lab or you don't.

DR. HARRINGTON: But as several people have pointed out, that's not necessarily -- we have no evidence that that actually changes your ultimate outcome.

DR. DOMANSKI: Yeah. I think that's fairly common practice, though, because one wants to know whether somebody has disease or not in order to treat them medically or with revascularization. It's not just a stent. It makes a big difference whether you have coronary disease or not in terms of how you treat somebody.

DR. HARRINGTON: Mori?

DR. KRANTZ: Yeah. I just had a question. I think -- it's a two-part deal. The first part, I think, it makes sense that they shifted midstream, and there's a lot of data that suggests that they had to reevaluate this, given the noise with adenosine.

I guess the second part, I do think it's less robust that when you go from a patient-level analysis to a population. I don't know what the statistical -- that's something I've always been taught, that, certainly, it's more robust when you look at patient-level versus population-level means. And I don't know if there's others that think the same.

DR. HARRINGTON: John?

DR. FLACK: Well, I don't have a problem with the normals being in there. I think if you want to know the performance characteristics of the test, you can't just test it in a high prevalence population. And this is the kind of population you're going to likely test it in. And so it makes sense to me. I think that they were justified in learning from the data and making a shift.

The problem I have with this is that you really need a more expansive adenosine/adenosine comparator. If you're going to basically make it an agreement-type study and say, we'll just take it at that, I just don't know if there's enough data there now, even though it kind of looks relatively similar, to be sure that it truly is.

I think that the end point that they went to I have some problems with. But I would agree with Emil and with Jim Neaton and some others that that endpoint, with other considerations, is not perfect, but acceptable.

DR. HARRINGTON: So you bring up an issue that several have brought up, the fact that the adenosine/adenosine comparison is only in the one study. And I think this starts to get to Jonathan's point about having relatively small numbers in that comparison.

DR. FLACK: Yes.

DR. HARRINGTON: That while it sort of all looks the same, you'd like to be a little more confident.

Is that a fair summary?

DR. FLACK: Yes, yes. And it's also fair to say I've had the cobwebs wiped from my brain and tutored. And I do understand where the bias is now. So sort of disregard my previous, stronger statements about the coronary angiograms.

But I think it would have been real helpful to have a more robust and larger adenosine/adenosine group because this notion of you've got error coming from other sources, and then you've got this test, sort of, as an adjunct to another test that has error and all, if you're really going to say that adenosine is acceptable in the gold standard, then you need a contemporary comparison with that that is convincing. It would be more reasonable, more convincing for me.

DR. HARRINGTON: Neil? I mean, Henry? Sorry.

DR. BLACK: Yeah. I'd just like to reiterate something that Sanjay said, and I think it's really very important. Regardless of what we do with this, if we're ultimately going to ask to approve it and put a compound on the market that will be better tolerated, how much sensitivity are we willing to sacrifice in order to do that? And I don't think we can tell from what we have.

There are just not enough comparisons. And 20 percent sensitivity, that's a lot for a screening test because that's what you want to use it for.

DR. HARRINGTON: I think that's in part getting to Lyle's point, that we didn't have enough look across all -- with different assumptions being made.

Go ahead, John.

DR. FLACK: Just one quick thing. I'm going to make a pitch that'll probably show I'm not an interventional cardiologist, that I think we've got to balance the missing of disease with the much larger numbers of people who will be taken down to the cath lab.

I've rounded on people who have had dye shot in them. They're not all low-risk people. They have low-risk histories, and they end up on dialysis, or they end up with problems and all. And it is not a benign thing to simply catch all the disease but then hurt a group of normals in the process.

I think what we have to do is we have to figure out where to balance these false positives and false negatives. And we may have a different sort of comfort level where we do that. But I think at the end of the day, it's got to be a balance and it can't just be simply, we've got to get as many of the positives as we can.

DR. HARRINGTON: No, I would fully agree that keeping truly normals out of the cath lab is a laudable goal that we don't want to bring people to the cath lab who we have some evidence which tells us that they're not going to benefit and may well be hurt by what we do in the cath lab. I fully agree with that statement.

Let me look around. Who hasn't had a chance yet? Neil?

DR. WEISSMAN: Most of my thoughts have been expressed. I don't have a big program with the mid-course correction. I think the idea that Frank said, the relative value of the original versus the revised, is reasonable.

I think the SDS is a clinically reasonable approach. I still think that although in clinical practice we compare stress to rest, what we really want to do here is identify the value of a new stress agent. You know what I mean? And we keep going back to this clinical way of doing it, stress to rest, that difference, stress to rest, that difference, the difference between those two differences.

I'm not sure. I'm a clinical cardiologist. I'm not a statistician. But I'm not sure that's what we're trying to get at here.

DR. HARRINGTON: Jim Neaton had brought this up several times today, about why just -- are you saying that the summed stress score to you might be a better indicator of whether or not something's comparable?

I think that's what you were getting at earlier today, Jim, is to --

DR. WEISSMAN: You know, look. We're looking at test/retest variability. That's acquisition variability and interpretation variability. We could measure interpretation variability. So then it's the acquisition variability. But here I think a large part of that variability is coming from the MPI. It's coming from the SPECT. And there's also acquisition variability that's introduced from the stress. I think we need to separate those two things out.

So in a way, a simple-minded way, not a

statistician way, if you look at the variability of the rest/rest in the same patient, and then the variability of the stress/stress in the same patient, I don't want the variability of the stress-stress to be any higher than the rest/rest.

DR. HARRINGTON: Okay. Peter?

DR. WEISSMAN: so I guess, to sum up, the robustness -- I have trouble withis it a robust measure because I have trouble defining that. But I think I'd have increased confidence if some of these other analyses got us to the same place.

DR. HARRINGTON: So you're falling into the camp started by Dr. Neaton that you don't fundamentally have a disagreement with the average summed difference score, but you want to see it in context with other things. And you're willing to consider the robustness of the overall data, as opposed to putting everything on sort of one single measure.

Is that a fair interpretation?

DR. WEISSMAN: Fair, yes.

DR. HARRINGTON: Peter?

DR. CONTI: Well, I kind of agree with Henry.

I feel waterboarded right now.

[Laughter.]

DR. CONTI: Get my attorney and get out of Gitmo.

One of the concerns I have in this whole thing -- I actually don't really understand what SDS means, you know, as a nuclear medicine physician and radiologist. I'm not really sure I get that, to be honest with you.

But that aside, I am very concerned about the number of patients that fall outside the normal category in each of these. If you look at all of these charts, you're talking about 50-some-odd patients in each of the situations, whether it's adenosine, calling it normal, and the binodenoson, calling it abnormal, or vice versa. It's always 50-plus patients.

That's a big chunk of patients, in my opinion, given the total number of patients that have been studied in these trials. I'd much rather see 2s and 3s and 34s and 35s and 27s and things like that. So I have a gut feeling that I don't like the way the study was done.

Having said that, they have done the best they can, I think, with the data that they have to perhaps repackage it and convince us that it is valuable. In my opinion, they ought to look, as I said earlier, at the 305 study, get a side-by-side, get some really baseline information about what adenosine versus adenosine can do, how it behaves; learn a lot about inter-reader variability that way, about the test variability that way, and then redesign the study with a new consultation with FDA.

DR. HARRINGTON: So you, too -- at least, an issue that's emerging is insufficient adenosine/adenosine data.

Dr. Fox, do you want to weigh in here?

DR. FOX: Yes. So apologies for hogging the microphone today.

[Laughter.]

DR. FOX: I think that the sponsor -- well, maybe first I'll make a comment about the agency. I think Dr. Rieves in particular in his opening comments made it very clear that the division has struggled with how to best evaluate these data, and hasn't just rejected them out of hand. So some credit to the agency for that.

Credit to the sponsor for, as some other people have said, trying to learn from the data along the way, carefully picking their way amongst the land mines of not doing something inappropriate, like trying to reanalyze data after unblinding and so forth.

I think that compared to many, many, many imaging studies in the literature, they've conducted the blinded reading and evaluations in a way that really adheres to what I can see as the highest standard.

It was kind of passed over quickly, but they took individual scans that had already been evaluated and kind to drop them in at random to the readers to assess any drift over time in the ability of the readers to adhere to something resembling objectivity.

Even though these are true tomographic techniques, the visual images, I think, display quite well. They're fuzzograms (ph), and nuclear medicine docs and echocardiographers and others working in imaging deal with the challenges of trying to come up with a meaningful clinical interpretation of these images every day in their work.

Still, I think it's worthwhile to mention that -- we've been debating efficacy here. And although it clearly has an influence on whether a patient gets taken to the cath lab or not, it doesn't in fact drive the decision in an absolute way. In the end, it's still a clinical decision up to the practicing physician, taking these laboratory data along with all the other data and making a determination.

Specific to the question at hand, I'm actually kind of unimpressed with the kappa statistic as a robust measure of anything, kind of as it being somewhat of a one-dimensional collapse of all of the data. And even though I agree with some of the comments made about the pluses and minuses of the SDS like the parameter, as I think Frank mentioned and Sebastian mentioned, the sponsor took, I think, quite a bit of effort to tease apart the various sources of variability that are inherent in this rather complex mix of pharmacologic agent, imaging test, clinical interpretation, and so forth.

So I guess I agree that, by itself, if this were a big outcomes trial of a pharmaceutical agent, if it didn't meet the -- if it didn't cross the line, it didn't cross the line. But I don't think that's the right analogy here. So I would say I would agree with sort of the Neaton approach, that by itself it may be not a standalone robust measure, but given all of the data so that you can understand what the measurement means, then it's probably reasonable.

DR. HARRINGTON: Jonathan, you brought up something that we hadn't commented yet that's probably worth commenting upon, is that while we may quibble with the design -- not enough adenosine/adenosine -- and we may quibble with the choice of endpoint, the sponsor and the steering committee for this study did a really careful job at what they set out to do, that the quality of the QA, et cetera, on the imaging was actually very well done.

DR. FOX: Yeah. I think that's really an important point to make. Just in my work, I've encountered core labs who will claim to be, you know, sort of practicing at a very high level of science when it comes to evaluating images; and when you ask them questions around, well, how often do you do a validation test, or where's your latest validation report, they say, what's that?

DR. HARRINGTON: Yes. So they should get kudos for that. This was core lab work that was very carefully done.

Jim?

DR. TATUM: So we're kind of drifting a little bit away from the question, so I figured I'd --

DR. HARRINGTON: Per usual.

DR. TATUM: Yes. So I agree pretty much with the idea that I'm not a fan of adenosine to begin with, and I think this is equally as bad. And I think they pretty much have proven that, and I think the change was appropriate.

I guess one of my big concerns goes back to where we started with Dr. Bengel, and that I understood there's no preclinical data to look at the reproducibility in the quantitation of what this drug does, particular with serial or different times over and over again, which could be done in a model. And from the institute I'm coming from, we've become nonclinical back to the bench very frequently. And I think this may be another place where we might want to go back and actually look at this, not only for this drug but for adenosine as well, which I think could be easily enough done.

The second thing that I think should possibly be -- no, let me go back to another point. We've talked a lot about let's incorporate a lot of clinical variables and everything else to it -- and Dwaine can comment on this -- but, realistically, we really don't do that most of the time in the regulatory arm. We need to have measurable things that are statistically looked at and those kinds of things. So I don't know the practicality of actually moving in that direction to do a trial, and I would not advise the sponsor on that. I believe that's between the FDA and the sponsor.

Another piece -- I'm trying to figure out all the parts of variability here. One I think is may be possibly a variability in the hyperemia. That's question one. That's fundamental. We kind of need to know the answer to that.

The second part is the analysis itself. And I'd like to see the data actually done with computerized data but doing it side by side so that basically these are actually merged, the variations decreased, and the analysis is run duplicated on each one without humans being involved, basically, after it's done. That would answer kind of another interesting question.

The other thing I'm concerned about is the broad range we saw on the one as far as what the perfusion reserve was. And I think, as Jim mentioned, one of the problems we have with most of the agents we use, there's a roll-off function. And at some of those higher-end things, I'm beginning to wonder we're losing the discrimination effect because of distraction problems that may be going on at the same time.

This would be a great PET trial. Rubidium, number one, would give you a perfect -- or if you want to use some of the other compounds, give you a perfect rest study because the rest study in most of these studies is not a very good statistical study. It's a low-dose study that's done in a way that's not really comparable. I never liked that very much, either.

But for this, it's perfect. It's a bolus injection. You get a prolonged enough time of hyperemia you could get the stress and you could get the rest. It's tomographic. It's attenuation- corrected. The whole bit. It just makes a lot of sense to trying to solve some of the problems that we're actually looking at here.

Then last, let me go back to the safety issue. The most common reason, I believe, for not doing a persantine or an adenosine is because somebody feels they have obstructive lung disease -- not even necessarily asthmatic, not even, you know, significant. And as was mentioned again, we go to dobutamine, which is a horrible stressor, in my opinion. So that I'd have to weigh into the picture as making this available to a group of patients right now that are unlikely to use it or get a clearly inferior stress test at the same time.

The other thing that's kind of interesting about this drug is that it does have a prolonged hyperemic phase. Adenosine chops off. Aminophylline chops off when you're using it with dipyridamole. There's a lot of data from that prolonged piece, particularly if you're doing a very rapid acquisition with motion at the same time. So again, it might be a nice fit for that.

Then the last thing I wanted to say, if this goes to approval, I would suggest a post-approval safety monitoring. And the reason I say that is there will be a perception of safety that could lead to utilization that is not exactly what you want and monitoring that you may not want, and maybe not the use of drugs for aminophylline, in particular, when you need it.

So I think that's important. And also, I don't think we have enough numbers, even though we have statistics on the safety for some other things that may occur like more angina, more infarctions, those kinds of things.

That's my whole list.

DR. HARRINGTON: Great.

So. Dr. Rieves, I think we've had a good discussion around this question. And maybe I can summarize the remarks into three major points, which I seem to be hearing over and over. And I'll look around for big disagreements here.

I don't think the panel has a problem with the changing of the analytical plan as more knowledge became available. In fact, I think many would say that that was a very reasonable thing to do as more knowledge was accumulated.

There is an issue that perhaps -- that I've heard from several people -- this as a single endpoint doesn't make it for a lot of the reasons that Dr. Levenson brought out, but that the panel is willing to look at that endpoint in connection with other analyses that might be supportive. So maybe not rising to the level of robust, but a reasonable measure, particularly combined with other things.

The second thing I seem to be hearing is that people are really -- because adenosine itself, as the reference standard, seems to have challenges, that there seems to be a consensus that there's just not enough adenosine/adenosine comparison in this package of information.

Then the third thing I seem to be hearing is that, in general, while people accept the premise of noninferiority, they're a bit troubled by -- or maybe to use Sanjay's word, they're not persuaded by the clinical margin that was set out.

Is that a fair summary as I look around the table?

[Affirmative nods.)]

DR. HARRINGTON: Okay. So we'll move to the next one. And some of the discussion we've already had, so some of this will go quicker.

DR. RIEVES: Well, that's actually what I was going to think. You largely answered question number 2 there.

DR. HARRINGTON: That's why I was going to quickly go through that. That's exactly what I was going to do.

So the second question, everyone, is that all three of the Phase 3 studies failed to achieve success upon the original primary endpoint of MPI concordance; however, success was achieved upon the revised endpoint.

Does this inconsistency impact your assessment of agreement between the two agents?

I think we've talked about that. If there's any -- okay.

Sanjay?

DR. KAUL: Well, they undershot their original primary endpoint by a considerable margin. But at the same time, they overshot their revised primary endpoint by a considerable margin. And in my mind I'm having difficulty reconciling which one is the right one and which one is not the right one. In other words, it has induced uncertainty in my mind. AAnd whenever there's uncertainty, it challenges the interpretability of the findings, and we have to go back to the first principles. We cannot eliminate uncertainty and attain certainty; we can reduce it by studying more.

So one way we can reduce this uncertainty is to study. And what I'm hearing here is that there seems to be a disagreement between what the agency considers to be an acceptable and valid design and endpoint and what the sponsor does. And perhaps that will allow them to sort of come up with a design and an endpoint where they converge on.

So that's what my recommendation would be, study more.

DR. HARRINGTON: Yes. You know, I actually wrote that note to myself earlier today, that this isn't a P equals .04 versus P equals .06, where you're sort of both hovering at the margin. I mean, they look really different when you analyze them in these two different ways.

So where is the middle ground? And I think this gets back to Jonathan's point a little while ago, is that there's just maybe not enough information here. And it's not necessarily that the particular test is a bad one, but there's just not enough of it.

Peter?

DR. CONTI: Again, I just want to point out that I think by going back, and even though it's not acceptable for Phase 3, to do that side-by-side, to see the consistency of the results across these three additional trials, would be very helpful.

They'd say, well, now we've eliminated some variables. We understand the data. We go back to the agency and we sit down and come up with a compromise as to how to do a follow-up study that would make sense, that will answer the specific questions, and in fact may be better to be done in different patient populations and not this composite of questionable disease, all the way up to known disease. I think we're struggling with that as well.

DR. HARRINGTON: Good comment.

Darren, we haven't heard from you on these two questions. You want to weigh in?

DR. McGUIRE: Well, I remain concerned about the level of discordance qualitatively between the two strategies. I think the criteria for approval referenced to a standard with a high level of agreement I think has two important criteria that need to be present.

First is that the referent is worth beating, and the challenge here, as the adenosine has performed so poorly, is we don't know -- even if it yielded identical results to adenosine, I'm not sure I would be any more or less convinced of the efficacy.

So one thing I'm concerned about in this specific field is, is it possible to do reference-based comparisons or do we have to go back to truth standards?

So the underlying concern I have is I'm afraid we lose the level of discordance when we go to the SDS delta endpoint. And again, as I said before, my interpretation of this endpoint assumes a certain acceptable level of concordance within the two diagnostic strategies before you can do the overarching population comparison. And I'm not convinced that that exists. I've gone back in context. I have Neil's comments about changing two groups. You know, we talked about earlier there was discordance by one or more groups in 26 to 34 percent; there's discordance by two or more groups, and the least is 11.5 percent and the most is 14 percent.

So those are still real numbers. That's 10 to 14 percent of patients that leave the cath lab with a completely different result, now differing by two qualitative severity classifications. And I'm surrounded by interventional cardiologists who believe the only reason to diagnosis coronary disease is to revascularize. But we actually have medications that we may prescribe in response to these studies.

So even a difference of 1 or 2 severity scores may prompt the prescription for aspirin, intensification of statin therapy, more intensive blood pressure reduction above and beyond. So again, that may be a truth standard to consider, whether it informs clinical decision-making.

But I'm convinced, looking at these data -- I'm optimistic that this agent will have a utility, and I'm optimistic that the safety profile is real, and its tolerability is superior. But I'm still left uncertain whether it's clinically relevant with regards to efficacy.

We have the risk, when comparing with an imperfect reference standard, of making a material step backward. And that's my greatest concern. And so, again, it's the underlying discordance from the raw data that leads me not to accept the SDS delta as the primary endpoint.

DR. HARRINGTON: And so you're also moving along with Sanjay, which is that when there's uncertainty, get bigger numbers --

DR. McGUIRE: More, yes.

DR. HARRINGTON: -- more data to try to limit what that uncertainty is.

DR. McGUIRE: Right. And I honestly believe we need truth standards, whether that's cardiac catheterization. The decision to go to the cath lab, in the cath lab the prevalence of obstructive disease, the ultimate revascularization, whatever that truth standard endpoint may be or, ultimately, clinical outcomes, which I don't think is -- I mean, that's a huge study.

But I think we have to define something that's clinically relevant to convince us that we're not stepping backwards clinically with regards to patient care and outcomes.

DR. KAUL: Can I make one follow-up comment to that?

DR. HARRINGTON: Absolutely.

DR. KAUL: I think what Darren said, a step backward, the potential for that in my mind is not inconsequential. It kind of reminds me of the bio creep. If we were to approve this drug, and because of its superior tolerability profile, it might conceivably become the comparator for future studies. And if there's any doubt about the efficacy with regards to the old standard, then I think there will be a significant bio creep in terms of efficacy. So the potential for that is real.

DR. HARRINGTON: Okay. I think, Dr. Rieves, probably the summary that we gave on the first point brings out much of what you wanted on this one as well.

Is there anything that -- okay.

So let's go to 3, which was, I thought, an interesting question regarding a discussion that we had had that John Flack had started about the angiography.

Knowledge of the MPI results may have impacted the decision to perform coronary arteriography in the Phase 3 study population. As we've heard, approximately 16 percent of the population underwent diagnostic cardiac cath.

How useful are those data from coronary arteriography images as the truth standard for establishing the binodenoson-based MPI performance characteristics?

So, John, you've commented. You've said you had the epiphany when you ate your cookie during the break, and you've --

DR. FLACK: No. My blind spot cleared. I don't think it's very useful and all. But in the future, it seems that, in an unbiased way, if you could have a comparator group out there or a group or subgroup that was sent in an unbiased manner without going through the filter of one test or the other and then be able to compare it, that would make sense.

Hopefully, at some point -- and this is beyond this study -- the FDA is really going to seriously look at the truth standard and maybe come up with some alternate, more contemporary endpoints to look at other than simply an anatomic one.

DR. HARRINGTON: And I thought you brought up an excellent point this morning. We know that not everyone with an abnormal SPECT has -- an abnormal MPI has obstructive coronary disease. We know that. And it's particularly an issue, perhaps, in women, particularly an issue, perhaps, in diabetics.

DR. FLACK: LVH, probably. Yes.

DR. HARRINGTON: Left ventricular hypertrophy. Long-standing hypertension. So there's other issues here.

But you also said something earlier today which I think gets at the essence here. You said that when you first looked at the angiography data, you found that compelling because those people had gone to the cath lab.

Get back to Darren's point here. If the sponsor chose to go out and do further study using the cath lab as a truth standard and designed a study, not where they would be selected to go but that's the way they went, would you find that level of evidence compelling?

DR. FLACK: I'd find it definitely more -- yes. I'd find it more compelling than what we have now. And given where we are, it's probably getting close to the best they're going to do outside of doing an outcomes study.

DR. HARRINGTON: Although, I think we had the discussion that we would think that the angiography data are flawed by the way they were done, that we still believe that angiography would be a nice way to match the test with coronary anatomy.

DR. FLACK: Yes. I think it's fairly reasonable to do that.

DR. HARRINGTON: Go ahead, Jim, and then

Mori, Emil.

DR. TATUM: I think if you're going to do an angiographic study, you need to consider Doppler. I think you need ultrasound. I think you need to look at flow reserve, because you can get sequential small lesions in vessels, and you can get significant hemodynamic flow with disturbances with vasodilators.

The other thing, the complexity of the anatomy is very important when it comes to vasodilators, steal phenomena being the one that really gives you ischemia. We're not touching any of that in what I think we're seeing here right now.

So I think if you're going to spend the money and you're going to do the effort to do this, you need to really do this correctly and get the truth standard you're really looking for.

DR. HARRINGTON: So that was John Flack's point, I think, earlier.

Let's go to Mori, Emil, then Peter.

DR. KRANTZ: I think what Jim's saying is accurate. But it's a big study; 1500 patients prospectively, most of them completely normal like me, and you're going to subject them to coronary arteriography and Doppler flow wire? So I think it's a daunting prospect.

I think another approach would be to retrospectively identify people who have had coronary arteriography, where you know their anatomy, and where you still have 50 or 70 percent stenosis in an epicardial vessel as a marker, and then go ahead and look at those folks. You might be able to power that with a smaller amount of patients.

DR. TATUM: If you're going to do it that way, that kind of technique, I think that makes a lot of sense.

DR. HARRINGTON: Neil?

DR. PAGANINI: You know, I guess I'm still -- this seems more like a study of the effectiveness of MPI, and that's not what we're here for. We're here to see whether or not the new drug is as good as adenosine. And I would agree that there has to be some sort of another standard to look for MPI. But let's just look at one versus the other. I don't think we have enough data on adenosine and its effectiveness. I think the 16 percent was backed into. It wasn't a prospective. It wasn't part of the study. It was backed into only when they unblinded for those that had adenosine, and then clinically they went on. So this is meaningless, as far as I'm concerned, for anything future.

So if you're going to do this -- and you're raising a larger question, I think. And the question is, is MPI that effective in what subgroups of patients for whatever? And I don't know if the data exists or not because I don't do this stuff. So if it exists, then just apply that data to these. But if it doesn't exist, then what you're doing is you're raising a much larger question than just one drug versus the other. But it's the test itself and the effectiveness of the test itself for either capturing those that should have caths or avoiding catheterization in those that shouldn't.

I think that's where we're talking. So your standard here is on quicksand, I think. It's sort of a morphing standard in morph world rather than the FDA.

DR. HARRINGTON: I think we had Peter down there. Then we'll go to Jonathan.

DR. CONTI: I agree on the cath side. I mean, the fact is that these should be probably patients that are destined to go to the cath lab as a requirement of the study, and then these other studies can be added to them. And this way it will give you -- it's a smaller study. It's a more directed study. It answers a more specific question, and you can move on.

As far as the MPI specifically, Jim brought up rubidium. I'm the director of the PET Center at USC. I've been trying to stay quiet. But the fact is is that that could be another way to approach this, where you do SPECT or PET as a follow-up truth, or just do the study directly in PET and avoid a lot of the technical issues.

DR. HARRINGTON: You mean some of the variability, et cetera, issues?

DR. CONTI: Yes.

DR. HARRINGTON: Jonathan?

DR. HALPERIN: Yeah. Just maybe a clarification. The way this is written, "Knowledge of MPI may have impacted the decision before arteriography," well, you know, it was not protocol-driven at all, as the sponsor pointed out. And in fact, the adenosine results were provided to the site as, okay, you ordered your clinical test; we grabbed your patient for our trial. We put them back. Here's your test. Go do what you want to do.

They did what they wanted to do. Okay? So the fact is that 16 percent of them underwent a procedure completely driven by the clinician on site.

As far as the second half, how useful, not very, for all of the reasons people have discussed. However, I was kind of impressed that you took people for whom the adenosine was, you know, this person should be cathed, and it was a coin flip in the end, and then you retrospectively, with all the weaknesses implied, apply the same question to the binodenoson results, and it was, guess what? A coin flip. So I wasn't surprised by that at all.

I think, to Mori's point, taking normal or low-risk people, all of them, to the cath lab, I'm not sure would be defensible. I don't want to use the big E word. But the idea of taking people with known lesions and then studying those and/or doing it by PET approach, I think all those would be valid.

DR. HARRINGTON: Good discussion.

Have you gotten what you need on this section, Dr. Rieves?

DR. RIEVES: Yes.

DR. HARRINGTON: So we're going to move now to the voting question. And before I do that, I'm required to read a statement prior to the voting procedure.

So we will be using the new electronic voting system for this meeting. Each of you have three voting buttons on your microphone, yes, no, and abstain. Once we begin the vote, please press the button that corresponds to your vote. After everyone has completed their vote, the vote will be locked in. The vote will then be displayed on the screen, and I will read the vote from the screen into the record.

Next, we will go around the room, and each individual who voted will state their name and vote into the record, as well as the reason why they voted as they did.

DR. KAUL: Can I ask a question, clarifying question? You know, uncertainty is relative. You have yes and you have no, but you don't have an option for "don't know." So abstain is the surrogate for that?

DR. HARRINGTON: Dr. Rieves, would you like to comment?

DR. RIEVES: We preferred the dichotomous outcome, candidly. We almost wish you would force a decision there. I would put "abstain in the extreme," if possible.

DR. HARRINGTON: Yes. We'd really -- you know, this is a difficult one in terms of there is a lot that we don't know here. But if you vote yes, you're essentially saying that you agree that the current data established a high likelihood of agreement between the two agents.

If you vote no, I think, Sanjay, that that would include don't know because the second part of that is, "Please discuss what additional data could be obtained," et cetera.

So abstain should be used rarely. I agree with Dr. Rieves. I think what's most helpful is if we vote yes/no and give our reasons.

Yes, Jim?

DR. TATUM: The other question I have is it says, "Do the Phase 3 study results establish high MPI agreement?" Are we talking everything that's been presented, the additional data we're talking about, or just the Phase 3 data and the outcomes that they've put forward originally?

DR. RIEVES: We're talking about the Phase 3 study results. And again, as we mentioned, we're looking at the totality of the results from the Phase 3 studies.

DR. TATUM: So everything that was presented today, in addition?

DR. RIEVES: That can be taken into consideration, right. That's part of the judgment, what we're asking you.

DR. HARRINGTON: Yes. The way I interpreted this -- correct me if I'm wrong, Dr. Rieves -- but if this was, did the primary endpoint make it, yes/no, you wouldn't need us. What you're asking is that based on everything we've heard today from the Phase 3, do we think as a group there is high agreement.

Is that fair?

DR. RIEVES: You're exactly right because, as we started the day out, we did not dismiss the product because it failed on primary endpoint. We don't want to commit a type 2 error, that sort of thing. So we want to give it the benefit. But we do want to force a decision, yes or no.

This is very tantamount to a risk/benefit type question, although we've made it a little bit more granular here. And the key question is how much sensitivity and specificity are we going to surrender, if you will, or we're in essence asking about those performance characteristics, but we must have -- and the key words are high agreement. And that's where the judgment comes in.

DR. HARRINGTON: Other questions? Sebastian?

DR. SCHNEEWEISS: Can we modify this to "reasonably high," or "what do you mean by high"?

[Laughter.]

DR. RIEVES: No, no, no, no.

DR. SCHNEEWEISS: "Clinically irrelevant

high"?

DR. RIEVES: That would not be useful. We do not need that advice. We need an answer, and --

DR. SCHNEEWEISS: But I think it's important for us, right, because, you know, we want to possibly weigh in the safety aspects as well in our comment, which is modified into reasonably high, and reasonably also with regard to what other evidence is out there.

DR. HARRINGTON: I think if you felt that -- reasonably high for you, if that constituted a yes or a no, you should vote that. But I think if we start putting qualifiers, we'll be here for a while.

Emil?

DR. PAGANINI: I need one more qualifier. I'm sorry about that. But this is -- as was said before, it's fuzzy data, the endpoint. The MPI is fuzzy. And you said, well, we do that in radiology all the time. So that's okay. I don't deal with shadows.

So the issue is, is this -- what you're asking, is this compound, within this fuzziness, the same as another compound that's in this fuzziness?

DR. RIEVES: Correct. And we're

not -- again, the threshold is one of high agreement. It's not, is it relatively similar or somewhat similar; do the data support the conclusion that there is high agreement?

Now, that's important not only for the reasons in terms of ultimate marketing of this product, but it also impacts -- we do have other products in development. It may have implications for the design of subsequent clinical studies.

DR. HARRINGTON: So, Emil, just maybe this will help you. I wrote down a few comments that Dr. Rieves made in his opening remarks, that, remember, that there's two ways, based on regulations, the performance characteristics and the agreement. And that's the one we're talking about here, the reference standard and new test. And he said this morning, I quote, that "The tests should be diagnostically interchangeable," and that, "High agreement is important."

DR. CONTI: I'm sorry again to ask another question, but there's high agreement with what we practice with in our daily activities and our experience with the adenosine; and then there's what was presented as part of the study adenosine. And we have talked a significant amount about what the adenosine data looks like is not necessarily being perfect or comparable to what our experience is.

So which adenosine are we comparing it to, their results or the general knowledge about how adenosine works in MPI?

DR. RIEVES: What is useful to us, all right, the charge to FDA is, FDA, do the data verify the claim? Do the data verify there is high agreement? It's not, is our gestalt, is our intuition, is our thinking that it is agreement when it's used in practice. It's do the available data -- hopefully our decision is going to be data-driven -- do the available data demonstrate high agreement?

DR. CONTI: Even to the point if the adenosine data was bastardized to be equivalent to the test drug, we'd have to go with that data.

Is that what you're saying?

DR. RIEVES: What I'm saying is there are multiple aspects that go into the consideration of robustness. For example, I don't want to go into the dialogue about the product we approved, for example, last year. But one of the major strengths of that database approval was that there was consistency and strong agreement on multiple types of outcomes from that. It wasn't just solely the primary endpoint, but there were multiple aspects that showed strong agreement. And the technical quality was assessed as appropriate.

DR. HARRINGTON: Okay. Last chance for questions.

Okay. I have one more statement I'm supposed to read.

Now that the discussion of the voting is complete, if there is no further discussion on the question, we will now begin the voting process. Please press the button on your microphone that corresponds to your vote.

[Pause.]

DR. HARRINGTON: Everyone has voted?

If everyone has voted, the vote is now complete and locked in. And now we're going to see -- so we have 15 voting yes, 11 voting no, and nobody voting to -- I'm sorry, 5 voting yes, 11 voting no, and zero voting to abstain.

Now that the vote is complete, we will go around the table and have everyone who voted state their name, their vote, and the reason they voted as they did. So why don't we start with you, Dr. Fox.

DR. FOX: Well, for some reason, the FDA put this duct tape on my buttons, so I -- no. I'm not a voting member, so I did not vote.

DR. HARRINGTON: Dr. Conti?

Sorry about that.

DR. CONTI: I voted no.

This is Peter Conti. I voted no because I felt that there was additional data that needed to be collected, and that what was presented, I think, was still insufficient to convince me that the drug is equivalent to adenosine at this point.

DR. WEISSMAN: This is Neil Weissman. I voted no because of some of the inconsistencies. I did struggle somewhat because I think that there is the possibility that more analysis of the data that exists

MI increase that confidence.

DR. HARRINGTON: Do you want to specify, Neil, what some of those analyses might be, at least in general terms? I think the FDA might find that helpful.

DR. WEISSMAN: I think it's the things that we talked about. It's trying to isolate out the variability from the MPI versus the stress, looking at segmental information, and so forth.

DR. FLACK: John Flack. I voted no. A single study, in all likelihood not enough people studied yet, and the adenosine/adenosine. And, really, I don't know that even with the differences that we see, if the bounds for noninferiority and all that or equivalence are well-said enough. I think they're on the right track, and they just need to accumulate a larger database.

DR. SCHNEEWEISS: Sebastian Schneeweiss. I voted yes because I felt I have to vote in the overall environment of great uncertainty in this field, the way I understand it, from the data presented today.

I certainly want to qualify that I would love to see more data according -- very similar to

Dr. Weissman -- the disentangle, where the variation comes from, the drug effect versus the imaging effect. But my answer has to be seen in the overall uncertainty of how this question is answered as of today.

DR. TATUM: Jim Tatum, and I voted no. And I think I did that based on my experience with the FDA, of knowing what identical, established, high agreement, and equivalency are. And those are high bars. They're not low. And they require real data to actually achieve those levels. And based on what we have, I think it's in the right direction. And again, I was kind of conflicted here on it as well. But if you look at those, at those outliers in particular on both of those, I just couldn't come to that level.

DR. BROMELING: I voted no because the --

DR. HARRINGTON: State your name.

DR. BROMELING: Lyle Bromeling -- because the kappa statistic never showed agreement, and there was a lack of power studies. Although there's probably enough power, I didn't see the power studies explicitly for the SDS type.

DR. KAUL: Sanjay Kaul. I voted no. I had issues with the design. A particular issue was lack of internal control in two out of the three. And we may debate the validity of the internal control, but I think that was one key element.

I was not able to interpret the endpoints and also establish the validity of their equivalence margins. And so that's the reason why I voted no.

DR. KRANTZ: My name is Mori Krantz. I actually voted yes. I think there's a lot of limitations, certainly, in the database that we've all addressed and talked about.

I do think that in my gut, my clinical gestalt is there is moderate discernment of ischemic burden with this agent. And certainly, as we mentioned earlier, I think further studies, particularly looking at patients with prior coronary arteriography, is warranted.

DR. PAGANINI: Emil Paganini. I voted yes. I voted basically because of the weighing the severity outcomes and the safety outcomes versus the outcome of the drug. I saw the variability of the standard that they used, and this drug was as variable as the standard.

It's obvious that we will need more data to understand not only the test itself but also the various drugs. However, as far as equivalency is concerned, I think it was equivalent.

DR. HARRINGTON: Robert Harrington. I voted no. I struggled a great deal with my vote here because I have a lot of enthusiasm for the data that they showed us. If the safety data is as it appears, this could be a step forward for the treatment of patients who are having these tests.

But I had enough uncertainty that I felt that more data is warranted, better setting of the margins; more adenosine/adenosine comparison internally, not just externally, to be able to put this agent into context. But I hope that the sponsor interprets the 11 to 5 not as a negative against the product but just as a limitation of the data that's available thus far.

DR. BLACK: Hi. This is Henry Black. I also voted no. I would have preferred to abstain, but I thought that was not courageous because I do have considerable uncertainty and a lot of faith in what we saw about the safety.

I wish I knew how much sensitivity, if any, we were sacrificing. I think if I had a good handle on that, I could probably be able to say whether it was worth it for the tolerability.

I think the additional studies we need have been well-described by others who do this. And I think may be a lot of the answers are already there, so it may not take that long.

I also want to echo what Bob said about not taking this as an indictment of the product. I think it's going to be a useful addition to what we do, but I don't think we're there yet.

DR. HALPERIN: Jon Halperin. I share Dr. Harrington's assessment. This was a difficult decision for me. I think the direction of the data are favorable. I believe it has the potential to be proven a superior compound for the indication.

However, the data are presently insufficient. I would like to see more data on segmental analyses showing comparable segmental defect interpretation with respect to the adenosine standard, or -- and I will say and/or -- angiographically-defined coronary disease. But I think it's really a matter of needing more data rather than the data themselves are negative. Thank you.

DR. DOMANSKI: Okay. Michael Domanski. I voted no. I didn't struggle, but I did feel sort of badly about it for a couple reasons. One is I think that the sponsors did a remarkably good job in many ways of executing the study that they actually did. It's a lot of smart people who did a really good series of analyses, number one.

Number two is I think it probably is a better tolerated drug, and I wouldn't be surprised to ultimately see it in the marketplace once effectiveness is demonstrated because I think the safety data are compelling.

I think that the effectiveness of this data were not, and I think if the adenosine is right, if it's right, then there are too many misses with this drug. But I'm not so sure which one is right. You know, the intriguing thought occurs to me that this drug may in fact be superior to adenosine. I'm not convinced it's the same, but it may also be better.

I think if there had been angiographic data, they might have won big with this one. So anyway, I hope it comes back and that it ultimately gets marketed.

DR. McGUIRE: Darren McGuire. I voted no. I'll pretty much just echo Dr. Domanski's comments. I also congratulate the sponsor for their rigor and the validity of the data that we've been presented. I have substantial optimism that this compound will have utility.

I remain unconvinced that we have comparable efficacy. I won't be surprised if it turns out to be superior to adenosine. But I think we have to rigorously assess that, given the magnitude of the problem.

So I think we do need a truth standard, and my preference would be cardiac catheterization as the truth standard.

DR. NEATON: Jim Neaton. I voted yes. I found it also a difficult decision. However, I felt that the omissions, in my own mind, were in the data set and were basically there that the -- between the sponsor and the FDA, they could resolve, and that there was a high likelihood that these two agents were similar to one another.

I attributed, you know, and may not -- it would be nice to have more data. But based on what we saw, the disagreements, you know, the relative disagreements between adenosine and the B drug is just chance.

And so that I agree with, you know, the statement made. You don't know whether -- it's true. You may not know whether adenosine is right or the B drug is right. But that's what you would expect, given the level of error, that you'd expect some in both directions. And that's what we saw. And so that's how I came to my conclusion.

DR. BENGEL: It seems for some reason the yes fraction has the last words. I'm Frank Bengel, and I also voted yes. And I'd like to -- I mean, most of my argument -- most of the arguments for -- most of the reasons why I voted yes have been brought up by the others already. But I'd like to make some more comments.

I think we did not discuss a therapeutic agent today, and we did not discuss myocardial perfusion imaging in general. We discussed a stress agent, and I think the data that were presented today, in this entire soup of -- not very clearly definable soup of myocardial perfusion imaging quantitation, the data that were presented today were not only just one analysis, it was multiple analyses, all of them having maybe a little bit of a problem. But the sum of all these analyses was good enough for me to say that probably both agents are agreeable.

DR. HARRINGTON: So, Dr. Rieves or Dr. Unger, any final comments for the panel, or questions?

DR. RIEVES: Thank you very much. We've all really struggled with this. And we also -- we have the same sentiment. This may prove to be a very effective product. But the information, the feedback, your perspective, is very useful.

Does anyone else have any comments or questions?

DR. UNGER: One thought might be worth bouncing off the committee members in terms of path forward is, since this is meant to take the place of exercise in people who can't, using exercise as a standard, we didn't discuss that at all.

Is that a viable approach, does anybody think? Anybody have thoughts?

DR. HARRINGTON: I would defer to people who think about this particular test all the time. But we heard some comments this morning, Ellis, from Dr. Udelson -- or this afternoon -- that they're really different. You don't get some of the physiologic changes, you get what exercise with these agents.

I don't know. I mean, yes, the nuclear folks around the table are saying -- shaking their head no.

DR. CONTI: I think it would add more variables that we don't need. And we certainly have too many of them now.

DR. HARRINGTON: Well, I want to thank the committee for their attention and their diligence today. And please travel safely on your way home.

[Whereupon, at 4:42 p.m., the meeting was concluded.]